농 학 박 사 학 위 논 문

RNA-seq based Transcriptome analysis on domestic animals under various experimental designs

다양한 실험 디자인으로부터 유래된 가축화 동물의 RNA 시퀀싱 기반의 전사체 분석

2017 년 2 월

서울대학교 대학원

농생명공학부 동물생명공학전공

박 원 철

RNA-seq based Transcriptome analysis on domestic animals under various experimental designs

By

Woncheoul Park

Supervisor: Professor Heebal Kim

Feb, 2017

Department of Agricultural Biotechnology

Seoul National University

다양한 실험 디자인으로부터 유래된 가축화 동물의 RNA 시퀀싱 기반의 전사체 분석

지도교수 김 희 발

이 논문을 농학박사 학위논문으로 제출함

2017 년 2 월

서울대학교 대학원
농생명공학부 동물생명공학전공
박 원 철

박원철의 농학박사 학위논문을 인준함

2017 년 2 월

위 원 장 한 재 용 (인)

부위원장 김 희 발 (인)

위 원 원 성 호 (인)

위 원 윤 숙 희 (인)

위 원 조 서 애 (인)

Abstract

RNA-seq based Transcriptome analysis on domestic animals under various experimental designs

Woncheoul Park

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Today, next-generation sequencing (NGS) technology can produce billions of nucleotides sequences in a single run. In addition, hundreds and thousands of papers in various research fields are published using NGS technology. The NGS technology is now the most powerful tool for the biological science and evolutionary science, and it produces more information than the whole information of the previous studies. RNA sequencing (RNA-seq) is a recent technique that appeared presently after next-generation sequencing (NGS) was invented. In the study of gene expression profiling, Transcriptome sequencing is most appropriate because it enables a profile of the whole transcriptome. A full view of a cellular transcriptional profile at a given

biological point and remarkably improve the power of RNA detection methods are provided by whole-transcriptome sequencing. In the application of NGS approach for RNA, several studies were successfully implemented. In the near future, every researcher will use NGS for RNA such as RNA-seq routinely, but transcriptome analysis doesn't come easy to them. Therefore, this thesis is mainly about researches using RNA-seq and additional DNA re-sequencing with NGS, from complex data that including expression information and additional evolutionary information of genes.

In chapter 1, the general background of NGS was summarized. The history of sequencing technologies and the classification of NGS methods were presented, more detail, the separation of NGS methods such as genomic and transcriptomic, which was listed. The character of RNA-seq was summarized. The history of sequencing and gene expression were presented, and comparison between RNA-seq and previous technologies, and an overview of RNA-seq analysis was presented. Evolution of domestic animals (Horse, Pig and Chicken) was introduced.

.In chapter 2, by using RNA-seq data in a Jeju native pigs and a Berkshire pig in three different tissues (fat, liver and muscle), significantly change of gene expression pattern of response to breed in each tissue was investigated. Jeju native pigs (JNP) have been adapted to an exotic natural environmental niche. They have been known to be resistant disease and have a good meat quality because of higher tenderness, juiciness, redness and brightness than those of Western breeds. In order to understand the molecular mechanisms of JNP specific phenotype, here I conducted comparative

transcriptomics study using RNA-seq technology. I compared transcriptome between JNP and Berkshire in three different tissues (fat, liver and muscle). I identified differential expressed genes (DEGs) of each tissue between the two breeds. Among the DEGs, I found that 26 genes were related to meat quality and body growth. Among those genes, *MPZ*, *AADAT*, *IGFN1* and *MYBPH* were up-regulated in JNP. Therefore, I suggest that JNP has different gene expression profile which related to meat quality and body growth compared to Berkshire.

In chapter 3, by using RNA-seq data in 9 chicken broiler of 3 different calcium intake condition, significantly change of gene expression pattern of response to calcium–stress in kidney tissue was investigated. Chicken (*Gallus gallus*) was first domesticated from a wild form called a red jungle fowl that still runs wild in most of Southeast Asia. After then, the grey jungle fowl (*G. sonneratii*) was hybridized, which was occurred probably about 8,000 years ago, and domestic chickens have been selected to have ideal economic traits such as a meat breed and a laying breed. Among these breeds, a meat breed such as a broiler is the most distributed in poultry industry. In addition, calcium is essential for normal cellular function and blood coagulation. However, it has a decisive effect on the hypocalcemia and the hypercalcemia if calcium intake was less or more than adequate calcium intake, which is related to body weight gain and stress such as hypertension. So, I experimented about the body weight gain and feed intake from 4 chicken broilers per calcium intakes (0.8, 1.0 and 1.2percent) and generated RNA-seq data from 10 broilers for gene expression profiling. As a result, I identified differentially expressed genes (DEGs) using cufflinks (128 DEGs between 0.8

and 1.0 percent, 141 DEGs between 0.8 and 1.2 percent and 103 DEGs between 1.0 and 1.2 percent), and also 12 DEGs were identified by edgeR. I identified that these DEGs were related to hypertension and blood pressure through the KEGG pathway enrichment, the co-occurrence and the protein/protein interaction (PPI) network analysis. Next, seven DEGs that were randomly chosen were validated by quantitative real-time PCR (qRT-PCR). In summary, the objective of this study was to investigate the influence of increasing calcium intake in broilers kidney. Therefore, I suggested that higher calcium intakes than adequate amount in broilers can cause the reduced body weight gain that was related to stress-induced disease such as hypertension.

In chapter 4, previous studies of horse RNA-seq were performed by mapping sequence reads to the reference genome during transcriptome analysis. However in this study, I focused on two main ideas. First, differentially expressed genes (DEGs) were identified by de novo–based analysis (DBA) in RNA-seq data from six Thoroughbreds before and after exercise, here-after referred to as "*de novo* unique differentially expressed genes" (DUDEG). Second, by integrating both conventional DEGs and genes identified as being selected for during domestication of Thoroughbred and Jeju pony from whole genome re-sequencing (WGS) data, we give a new concept to the definition of DEG. I identified 1,034 and 567 DUDEGs in skeletal muscle and blood, respectively. DUDEGs in skeletal muscle were significantly related to exercise-induced stress biological process gene ontology (BP-GO) terms: 'immune system process'; 'response to stimulus'; and, 'death' and a KEGG pathways: 'JAK-STAT signaling pathway'; 'MAPK

signaling pathway'; 'regulation of actin cytoskeleton'; and, 'p53 signaling pathway'. In addition, I found *TIMELESS*, *EIF4A3* and *ZNF592* in blood and *CHMP4C* and *FOXO3* in skeletal muscle, to be in common between DUDEGs and selected genes identified by evolutionary statistics such as F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH). Moreover, in Thoroughbreds, three out of five genes (*CHMP4C*, *EIF4A3* and *FOXO3*) related to exercise response showed relatively low nucleotide diversity compared to the Jeju pony. DUDEGs are not only conceptually new DEGs that cannot be attained from reference-based analysis (RBA) but also supports previous RBA results related to exercise in Thoroughbred. In summary, three exercise related genes which were selected for during domestication in the evolutionary history of Thoroughbred were identified as conceptually new DEGs in this study.

In chapter 5, in this study, I aim to identity that differentially expressed isoforms (DEIs), differential splicing and alternative splicing event by using the published Thoroughbred racing horse RNA-seq data between before and after exercise, because previous studies didn't researched that carefully and without researches about alternative splicing event in Thoroughbred racing horses. I used g/--GTF-guide option in Cufflinks program, because I want to find the all reference transcripts as well as any novel genes, isoform and splicing. As results, In DEIs, the number of DEI in blood and skeletal muscle were 67 and 1,133 respectively. Among them, novel DEIs were 37 in blood, 378 in skeletal muscle. In addition, I identified 7 (6 up-regulated and 1 down-regulated) DEIs in blood and 56 (45 up-regulated and 11 down-regulated) DEIs in skeletal muscle. Among them, in blood, 3 isoforms such as *HSPA8*

(heat shock 70 kDa protein 8 gene), *RhoB* (Rho-related GTP-binding protein) and *SOCS3* (suppressor of cytokine signaling 3 mRNA) (up-regulated) in blood and 5 isoforms such as *AMPD2* (AMP Deaminase Isoform L), *ICAM1* (intercellular adhesion molecule 1), *MMP-1* (Matrix metalloproteinase-1), *MXD1* (MAX Dimerization Protein 1) and *TET2* in skeletal muscle were revealed that related to exercise-induces. Moreover, we identified 4 (4 up-regulated) significant differential splicing such as *BLZF1, ITGB6, KDM5C* and *ZNF207* gene in skeletal muscle. Most of these genes were included a litter-related exercise-induce stress with alternative splicing. Conclusively, we classified and identified the alternative splicing events in blood and skeletal muscle in six Thoroughbreds racing horses before and after exercise. As a result, we identified that exon skipping/inclusion (ESI) type is the most common of alternative splicing event, this is the identical result such as human and yeast but the different result as pig with alternative 3' splicing (A3)

Through these studies, diverse applications, of the transcriptome analysis considering the experimental design and purpose, was successfully demonstrated in RNA-seq data or additional re-sequencing data derived from NGS. By using data acquired from RNA-seq or additional re-sequencing technology, lots of biological and evolutionary meaning could be achieved. Given these results, I suggest that researchers in transcriptome study field will employ the suitable transcriptome analysis corresponding to their experimental design and purpose.

Key words: Next generation sequencing (NGS), RNA-seq, Domestic animal, Transcriptome analysis, differentially expressed genes (DEGs), *de novo*

assembly, differentially expressed isoforms (DEIs), splicing and alternative splicing event

Student number: 2009-21249

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1 Next generation sequencing (NGS)

1.1.1 History of sequencing technologies

Since 1940s, DNA sequencing technology has progressively developed. Two of the Nobel Prizes for Chemistry were given to British biochemist, Frederick Sanger, who developed this technology and is considered the backbone of this field. In 1955, Sanger and corroborators published paper that unraveled all amino acids of insulin for the first time (Ryle et al. 1955, Sanger et al. 1955). In 1975, Sanger and Coulson published "Plus and Minus" methodological paper for DNA sequencing (Sanger et al. 1975). 2 years later, they published two papers which included the method of the rapid determination of DNA sequencing and introduced the "dideoxy method" (Sanger 1977, Sanger et al. 1977). This "dideoxy method" provided the solution for the limitation of "Plus and minus method". In the same year, Maxam and Gilbert published the method paper that was more improved than earlier DNA sequencing technology and used the nucleotide sequence of a terminally labeled DNA molecule and reactions that cleave DNA Preferentially at guanines (G) , at adenines (A), at cytosines (C) and thymines (T) equally, and at cytosines alone (Maxam et al. 1977).

In 1986, the first automatic DNA sequencing was introduced by Applied Biosystems (ABI), for which different fluorescently end-labelled primers were used in each of the four didecxy sequencing reactions in Sanger's own "dideoxy method". The sequence could be detected by using a fluorophore

covalently attached to the oligonucleotide primer used in enzymatic DNA sequence analysis and the characteristic fluorescence spectrum is used for each of the reactions specific for the four bases such as A, C, G and T (Smith et al. 1985). They used the computer programs that automatically converted fluorescence data into a sequencing data without using autoradiography. Before long Smith's own automated DNA sequencing, Craig Venter and his colleagues at the National Institutes of Health (NIH) set up the six automated sequencers which was expanded to 30 sequencers in 1992 at The Institute for Genomic Research (TIGR) and in 1993 the Wellcome Trust Sanger Institute, was established (Adams et al. 1994). Using the automated sequencing, in 1991 Adams and collaborators developed the expressed sequence tag (EST) and discovered 337 new and 48 homolog-bearing human genes via EST approach (Adams et al. 1991). Long before, above 87,000 human transcripts fragments were sequenced by using EST approach. Nowadays, over 74 million ESTs from over 2,400 different organisms are available in EST database (dbEST) (Boguski et al. 1993)

In 1995, the first cellular genomes were the complete nucleotide sequence (1,830,137 base pairs and 580,070 base pairs) of the bacterium *Haemophilus influenzae* (Fleischmann et al. 1995) and the Mycoplasma genitalium (Fraser et al. 1995), which were sequenced by using whole genome shotgun (WGS) sequencing and TIGR assembler (Sutton et al. 1995) at TIGR. WGS was already used in 1979 for small (4000- to 7000-base-pair) genomes (Staden 1979), in which genomic DNA is sheared into random fragments, size-selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The major problems to assembling such projects: the large

number of pairwise comparisons required, the presence of repeat regions, chimeras introduced in the cloning process, and sequencing errors were solved by using the TIGR Assembler. In addition, using these approach, more complex genomes such as the ones of the yeast *Saccharomyces cerivisiae* (Goffeau et al. 1996) ,the bacteria Escherichia coli (Blattner et al. 1997), the nematode *Caenorhabditis elegans* (Consortium 1998) and the fruit fly *Drosophila melanogaster* (Adams et al. 2000) were sequenced.

In 2001 and 2003, the public human genomes were published by the group of organizations including the US government. Firstly, the idea of Human genome project (HGP) was created by the US government in 1984, then the project was formally launched in 1990 and finally completed in 2003.with the funding form the US government through NIH and many cooperators from around the world. Secondly, a parallel project was conducted by the Celera Corporation, which was formally launched in 1998. Then, published human genomes were published (Lander et al. 2001, Venter et al. 2001). These human genomes were made from WGS strategy and were caused a great sensation. After this accomplishment, the development of sequencing technology was accelerated. Particularly, a new generation of non-Sanger-based sequencing technologies had been created. With the unprecedented speed (Schuster 2007), the sequencing cost was gradually decreased while the number of bases sequenced was gradually increased (Mardis 2008).

1.1.2 Method and Classification of NGS

So far, Next-generation sequencing platforms is available in many and vary expanded ways, which provides the chance that researcher in a variety of areas is able to faster and deeper studies in it than first generation sequencing platforms. Sequencing methods is some different, first, how the DNA or RNA samples are obtained. Second, analysis options in used data. After the sequencing libraries are prepared, most of NGS platforms are similar and same. And NGS is used mostly in approximatively two fields (Genomics and Transcriptomics).

1.1.2.1 Genomics

Whole-Genome Sequencing

WGS or, full genome sequencing finds out the complete DNA sequence of genome in all organism at once. This includes sequencing of organism's chromosome and the mitochondrial DNA as well as chloroplast in plants. It provides the most broad collection of genetic variation such as rare variants and structural variants in an individual. In the initial stage, Genome-wide association studies (GWAS) is the microarray-based and have been the most common approach for identifying the genetic basis of traits and disease associations by the case-control setup which compares two large groups of individual's whole genome, one healthy control group and one case group affected by a disease. And the microarray-based GWAS can interrogate over 4 million markers per sample. However, the microarray-based GWAS usually explain only a minor fraction of the genetic risk and cannot account for the two additional factors such as rare variants and copy number variants (CNVs),

which can have important influences on disease phenotypes (Cohen et al. 2004, Estivill et al. 2007). Along with the lowering of the price of sequencing technology and achieving rapid production of large volumes of data, sequencing has become a powerful tool for genomics research, GWAS have been shifted from microarray-based genotyping studies to WGS. In addition, the WGS-based GWAS can interrogate 3.2 billion bases of the human genome. This flexible nature of the technology makes it equally useful for sequencing any species such as livestock animals, plant or microbial genome. Currently, there are many public and private companies for commercialization of WGS, a representative companies are Illumina (http://www.illumina.com/), Oxford Nanopore Technologies (https://www.nanoporetech.com/), Pacific Biosciences (http://www.pacb.com/), Knome (https://tutegenomics.com/), Sequenom (https://www.sequenom.com/), Complete Genomics (http://www.completegenomics.com/), Affymetrix (http://www.affymetrix.com/estore/), IBM (http://www.ibm.com/us-en/), Life Technologies (https://www.thermofisher.com/) and 454 Life Sciences (http://454.com/). These companies are supported by a vast amount of finance from venture capitalists, hedge funds and investment banks.

Exome Sequencing

Also known as whole exome sequencing (WES or WXS), is laboratory process for sequencing all the expressed genes in genome (known as exome). The human genome composed three billion DNA base pairs, while diploid genomes have twice the DNA content. However only a 1.5 percentage of

those nucleotides are actually translated into proteins (known as exons : expressed region) that estimated 180,000 exons, 20,000~25,000 protein-coding genes and the rest is associated with non-coding RNA molecules, regulatory DNA sequences, LINEs, SINEs, introns (intragenic region), and sequences for which function has not been defined as yet (Lander et al. 2001, Ng et al. 2009). The exome means the entire exons of a genome which are the coding parts of genes. The aim of WES is to identify disease-causing variants that is responsible for both Mendelian and common diseases such as Miller syndrome and Alzheimer's disease and can identify rare mutations that GWAS cannot determine. In addition, WES is a cost-effective alternative to WGS, which selectively captured the protein-coding region of the genome and sequence them. It can efficiently identify variants across a wide range of applications, including population genetics, genetic disease, and cancer studies However, WES has the some limitations that it is only able to identify those variants found in the coding region of genes which affect protein function, except the structural and non-coding variants. So WES can be used in a lot in fixed-cost researches, because that can sequence samples to much higher depth than what can be achieved with WGS. However, with the reduction of WGS costs, that technique it will eventually replace all WES because it offers a view at all regions of the genome, not just those that codes for proteins (Gilbert 1978).

***De Novo* Sequencing**

The primary generation sequencing of a particular and novel organism such as viruses, bacteria, lower eukaryotes, higher eukaryotes and etc is called de novo sequencing that was accomplished by capillary electrophoresis (CE) sequencers that made overlap consensus assembly the gold-standard technology for de novo projects, with its long read lengths and high accuracy. In addition, pyrosequencing was applied to generate the 500-kb genome sequence of Mycoplasma genitalium (Margulies et al. 2005) and hybrid approaches combining Sanger and pyrosequencing have also been proposed (Goldberg et al. 2006). However, more recently, the development of short-read assemblers and the high-throughput abilities of massively parallel sequencing (high-throughput sequencing) have significantly decreased the time and cost associated with a whole genome sequencing. For example, hierarchical shotgun sequencing approaches; the SHort Read Assembly Protocol, or SHARP (Sundquist et al. 2007). In addition, de novo sequencing has been used when there is no reference genome sequence available for alignment, for instance, bacteria, soybean, giant panda, human, Neanderthal and mammoth (Miller et al. 2008, Reinhardt et al. 2009, Green et al. 2010, Kim et al. 2010, Li et al. 2010, Qi et al. 2010). A detailed genome analysis of any organism is only achievable after de novo sequencing has been implemented. Sequence reads are assembled as contigs and the coverage quality of de novo sequence data depends on the size and continuity of the contigs (ie, the number of gaps in the data). Another important factor in generating high quality de novo sequences is the diversity of insert sizes included in the library such as shotgun and long jumping distance libraries. Combining short-insert paired-end in shotgun library and long-insert mate

pair sequences in long jumping distance library are the most powerful approach to obtain high quality genome. The combination of insert sizes is possible for detecting more wide range of structural variant types and is essential for accurately identifying more complex rearrangements. The short-insert reads that were sequenced at higher depths can fill in gaps, while the long inserts reads cannot cover those gaps, because sequencing with longer reads are done at lower read depths. Therefore, it is better to use a combined approach results in higher quality assemblies. In parallel with NGS technology developments, many algorithmic developments have been made in sequence assemblers for short-read data. Researchers can implement high quality de novo assembly using NGS reads and free short-read assembly programs. Moreover, these free short-read assembly programs can be implemented by existing computer resources in the laboratory, For example, the E. coli genome can be assembled in as little as 15 minutes using a 32‑bit Windows desktop computer with 32 GB of RAM. *De novo* sequencing was divided into two sequencing such as SOLID and Sanger, the benefits of SOLID sequencing include excellent coverage, greater confidence then Sanger and flexibility in the assembler you choose. The benefits of Sanger sequencing include long scaffolds, accurate and reliable detection, automated approach and reproducibility.

Targeted Sequencing

Targeted DNA sequencing is called “capture” technologies, it is a powerful and cost-effective way to detect and discover known and novel variants in

specific areas of interest such as selected subset of genes or regions of the genome that are isolated and sequenced. An early example of “capturing” specific areas of interest for next generation re-sequencing was the work of Thomas et al. (2006). In addition, targeted sequencing enables sequencing at much higher coverage levels. To generate a full view of related targets, you need complete and uniform coverage across specific areas of interest. Unfortunately, data in short-read sequencing is inclined to break downs and miss-mapping to extend repeats. Additionally, bias that were related to PCR can engender insufficient coverage for variant calling in GC-rich regions (Carneiro et al. 2012). For example, a typical WGS study accomplishes 30~50x coverage levels per genome, while a targeted resequencing can easily accomplishes 500~1000x or higher at the target region. This higher coverage allows researchers to identify rare variants that would be too rare and too expensive to identify with WGS or CE-based sequencing. Targeted gene sequencing panels are useful tools for analyzing specific mutations in a given sample. Focused panels contain a selected subset of genes or genic regions that have known or suspected associations with the disease or phenotype under various studies. Gene panels can be purchased with fixed, preselected content or custom designed to include genomic regions of interest. With such choices, researchers can target regions of the genome relevant to their specific research interests. Custom targeted sequencing is ideal for examining genes in specific pathways, or for follow-up studies from GWAS or WGS. Targeted sequencing comes in two main methods such as target enrichment and amplicon sequencing. With target enrichment have some benefits such as larger gene content (typically > 50 genes), more comprehensive profiling for

all variant types and more comprehensive method, but with longer hands-on time and turnaround time. Also amplicon sequencing have benefits such as smaller gene content (typically < 50 genes), ideal for analyzing single nucleotide variants and insertions/deletions (indels) affordable and simple workflow. This is particularly useful for the discovery of rare somatic mutations in complex samples (eg, cancerous tumors mixed with germline DNA) (Lo et al. 2009, McEllistrem 2009). Another common amplicon application is sequencing of the bacterial 16S rRNA gene from multiple species, a widely used method for phylogeny and taxonomy studies, particularly in diverse metagenomic samples (Ram et al. 2011).

1.1.2.2 Transcriptomics

Total RNA sequencing and mRNA sequencing

In the study of gene expression profiling, Transcriptome sequencing is most appropriate because it visualizes the profile of the whole transcriptome. A full view of a cellular transcriptional profile at a given biological point and remarkable improvement of the power of RNA detection methods are provided by whole-transcriptome sequencing. As with any whole-transcriptome sequencing method, almost limitless dynamic range enables identification and quantification of both common and rare transcripts. In addition, this method has an ability to align sequencing reads across splice junctions, as well as to detect isoforms, novel transcripts and gene fusions. Library preparation kits that support specific detection of strand orientation are available for both total RNA-Seq and mRNA-Seq methods. However,

total RNA-seq theoretically should detect more lncRNAs due to its RNA selection which is independent on the poly (A) tail. In addition, total RNA-seq costs more than mRNA sequencing (mRNA $500 versus total RNA $650) (Guo et al. 2015).

Total RNA sequencing (RNA-Seq) captures a broader range of gene expression changes and enables the detection of known and novel features of the transcriptome in both coding and multiple forms of non-coding RNA. Total RNA-Seq can accurately measure gene and transcript abundance. Total RNA-Seq provides optimal coverage in normal or low-quality samples. The Total RNA Sample Preparation Kit efficiently removes ribosomal RNA and other high abundance transcripts using Epicentre's proven Ribo-Zero™ ribosomal RNA reduction chemistry, with an improved workflow optimized for high-throughput studies. The resulting combination of high-quality ribosomal removal and sample preparation chemistries into a single, streamlined solution provides opportunities for researchers to conduct highly-accurate gene expression studies. Total RNA-seq have some benefits. It captures both known and novel features and allows researchers to identify biomarkers across the broadest range of transcripts. Moreover, it enables a more comprehensive understanding of phenotypes of interest and it allows profiling of the transcriptome across a wide dynamic range

mRNA sequencing (mRNA-Seq) has rapidly become the method of choice for analyzing the transcriptomes of disease states, of biological processes and across a wide range of study designs. In addition to being a highly sensitive and accurate means of quantifying gene expression, mRNA-Seq can identify known and novel transcript isoforms, gene fusions and other

features as well as allele-specific expression. mRNA-Seq delivers a complete view of the coding transcriptome that is not restricted by the filter of prior knowledge. The TruSeq Stranded mRNA Library Prep Kit offers a streamlined, cost-efficient, and scalable solution for coding transcriptome analysis. It is compatible with a wide range of samples. mRNA-Seq have some benefits. It offers a broader dynamic range, enabling more sensitive and accurate measurement of fold changes in gene expression and it captures both known and novel features. It can also be applied across a wide range of species.

Targeted RNA sequencing

Targeted RNA sequencing enables measuring genes and transcripts of interest for differential and allele-specific expression, as well as detection of a coding variant (cSNV) gene-fusion, isoforms, splice junctions and alternative splicing. In addition, it is a powerful method for the investigation of specific pathways of interest or for the validation of gene expression microarray or whole transcriptome sequencing results. Moreover, it can overcome the challenge that the wide dynamic range of the cellular RNA population by focusing on sequencing selected set of genes or genomic regions, so that the sequencing can provide huge read coverage (Levin et al. 2009, Mercer et al. 2012). This is possible to discover more precise gene or transcript, as well as quantification and assembly of even very low expressed transcript. Furthermore, in combination with multiplex library preparation, the increased efficiency of targeted RNA-seq can also reduce reagent costs. Moreover, targeted RNA-seq

could easily interrogate a variety of gene targets, but a proper member of samples by genes evaluated throughput calculation is about 100 sample by 100 genes. Below that number of samples and genes, straightforward real-time PCR approach would rather make sense.

Small, noncoding RNA and microRNA sequencing

Small, noncoding RNA, or microRNA s are short, 18~25 nucleotides that play a role in the regulation of gene expression often as gene repressors or silencers. Small RNA sequencing (RNA-Seq) is a technique to isolate and sequence small RNA species, such as microRNA (miRNA), pre-microRNAs, short-interfering RNA (siRNA), transfer ribonucleic acid (tRNAs) and small nuclear ribonucleic acid (snRNAs), small nucleolar ribonucleic acid (snoRNAs) and piwi-interacting RNA (piRNA). This technology is dependent on excision of a custom size for library construction (within 30-200 nucleotides), which can query thousands of small RNA and miRNA sequences with unprecedented sensitivity and dynamic range. With small RNA-Seq, novel miRNAs and other small noncoding RNAs can be discovered, and the differential expression of all small RNAs in any sample can be examined. This method can also help us understand how post-transcriptional regulation contributes to a phenotype. MicroRNA sequencing (miRNA-seq) differs from other forms of RNA-seq in that its input material is often enriched for small RNAs and it is dependent on automated gel extraction of a band representing insert size of 15-30 nucleotides. In length; this assures that contamination of degraded RNA, empty adaptors or primer dimers is minimal

in the NGS microRNA library. The study of microRNAs has grown as their role in transcriptional and translational regulation has become more evident (Dior et al. 2014, Wang et al. 2014). The microRNA library read depth is one of the most crucial factors with regards to both differential expression analysis and discovery of novel microRNAs (Metpally et al. 2013).

1.2 RNA sequencing or whole transcriptome shotgun sequencing

1.2.1 History of sequencing and gene expression

RNA sequencing (RNA-seq) is a recent technique that appeared after next-generation sequencing (NGS) was invented. Various technologies have been developed to deduce and quantify the transcriptome before RNA-seq. First, early efforts to explore transcriptomes used expressed sequence tags (EST). The EST technique refer to the creation of cloned cDNA molecules from mRNA templates and sequencing 3′ or 5′ ends using Sanger sequencing. (Sanger et al. 1991). Gene discovery in many species was catalyzed by this technique (Adams et al. 1993, Hillier et al. 1996, Marra et al. 1999). However, ESTs were not best for gene expression profiling, mainly due to their significant cost.

The development of short (14–21 base pair) serial analysis of gene expression (SAGE) tags and derivative technologies often deal with cost issues by making it possible to detect 30 or more number of expressed transcripts in a single pass sequencing read, as opposed to a single transcript

as in the EST technique. (Velculescu et al. 1995). This technique is useful for gene expression profiling because of the increased number of transcripts that was detected by SAGE (Polyak et al. 2001, Yamamoto et al. 2001). Since then, the development of the cap analysis of gene expression (CAGE) technique detected the bias of SAGE tags to the 3′ ends of transcripts (Kodzius et al. 2006). After high throughput sequencing (HTS) approaches, a new version of SAGE called 'DeepSAGE' was developed on the 454 sequencer, (Nielsen et al. 2006), then Tag-Seq that can be used to measure expression values of genes and with strand specificity derived from SAGE (Morrissy et al. 2009). Expressed sequences, longer than short tags, have also been analyzed using next generation sequencing. By capturing poly (A) + mRNA molecules and using a shotgun style approach akin to that previously defined for the genome, the entire mRNA content of a sample can be sequenced. This approach is known as whole transcriptome shotgun sequencing (WTSS) or RNA-Seq. RNA-Seq was used for a variety of applications in many research areas. For examples, the transcriptome of a human prostate cancer cell line, the HeLa S3 cell line, ovarian cancer samples, granulosa-cell tumors, large B-cell lymphomas and lymphomas cell line (Bainbridge et al. 2006, Morin et al. 2008, Shah et al. 2009, Shah et al. 2009, Morin et al. 2010, Steidl et al. 2011). Moreover, according to an ISI Web of Knowledge search in July 2015, the first publications containing the keyword "RNA-sequencing" appeared in 2008 and since close to 7,000 manuscripts containing this keyword have been published.

1.2.2 Comparisons between RNA-seq and previous technologies

RNA-seq has now mostly superseded previous technologies for transcriptome analysis, because of many reasons. First of all, RNA-Seq is not dependent on prior sequence knowledge, i.e., it can be applied to any system from which RNA can be isolated in sufficient quality and quantity. In contrast, the design of microarrays depends on prior sequence information, be it from genome sequencing or sequencing of expressed-sequence tags (ESTs). Secondly, RNA-seq provides a direct measure of RNA abundance in contrast to microarrays, which provide relative fluorescence intensities. Hence it is rather difficult to compare the results of microarrays between labs whereas this is more straightforward with RNA-seq data. Third, RNA-seq enables simultaneous sequence discovery and quantitation. Fourth, RNA-seq provides at least two orders of magnitude larger dynamic range than microarrays, which allows for the quantitation of low abundance transcripts in the presence of highly abundant transcripts, given sufficient depth of sequencing. Fifth, RNA-seq allows for the detection of sequence variants, which enables analysis of allele-specific expression in heterozygous individuals and the detection of sequence variants between individuals. Sixth, recent instruments enable highly multiplexed sequencing of hundreds of bar-coded RNA-seq samples in a single run, which makes RNA-seq relatively economic. In addition, two studies that transcriptome analysis in maize and Arabidopsis benchmarked RNA-seq data against previous EST and microarray work and concluded that transcriptome analysis by sequencing methods will soon replace these previous technologies (Emrich et al. 2007, Weber et al. 2007).

1.2.3 RNA-seq applications

In the application of NGS approach for RNA, several studies were successfully implemented. First, differentially expressed genes (DEGs) detection in conditions of interest (Robinson et al. 2010). Second, synonymous and non-synonymous variants identification (Lu et al. 2010). Third, transcript annotation based on reference genome (Roberts et al. 2011). Forth, novel transcript detecting that including exon, isoform and gene (Grabherr et al. 2011). Fifth, orthologous gene discovering among different species (Zhu et al. 2014). Sixth, de novo assembly when without reference genome and using the unaligned read (Xie et al. 2014, Kazemian et al. 2015). In many applications, the most population study is to detect DEGs using RNA-seq data

1.2.4 RNA-seq data analysis

RNA-seq has a wide range of application, but there is no optimal pipeline for a wide variety of different applications and analysis scenarios in which RNA-seq can be used. So I review some of the major steps in RNA-seq data analysis, including quality control of raw reads, read alignment, transcriptome profiling, differential expression and *de-novo* assembly. Figure 1.1 illustrates a generic roadmap and commonly used strategies for experimental design and analysis using RNA-seq with standard Illumina sequencing (Conesa et al. 2016).

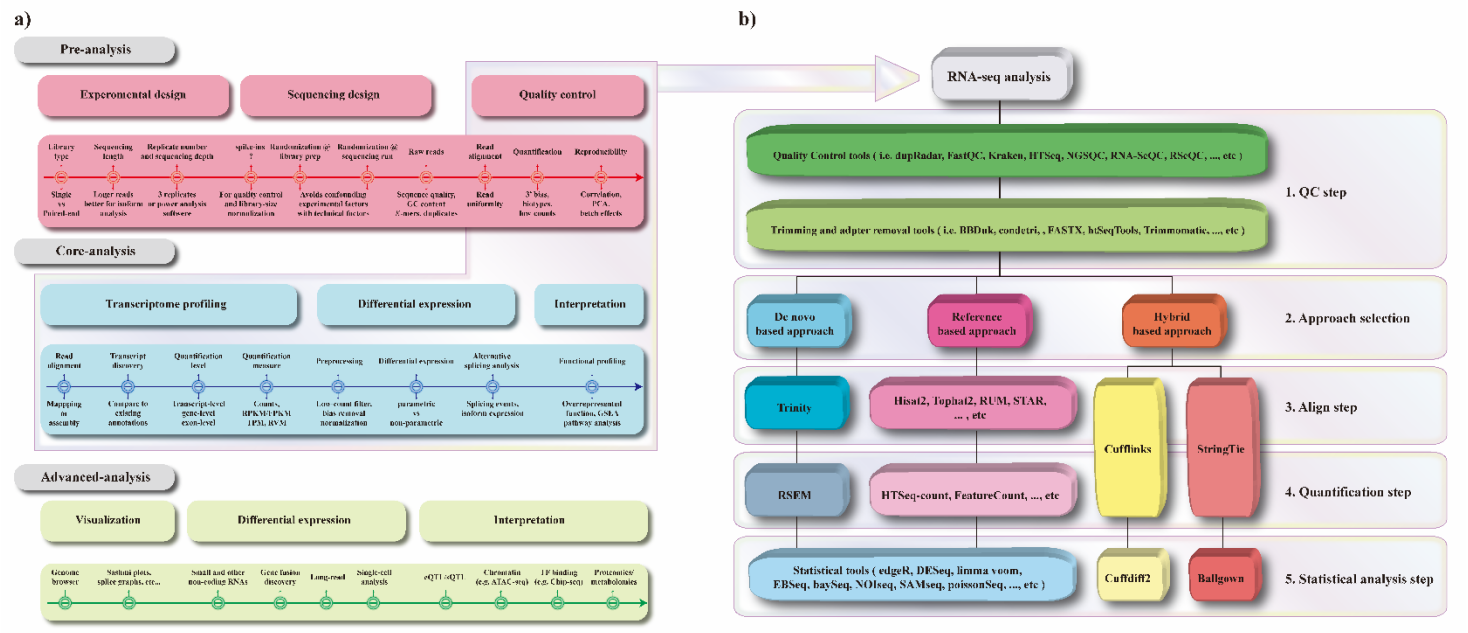


Figure 1.1. A generic roadmap for RNA-seq computational analysis. **a)** The major analysis are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **b)** Commonly used strategies for regular RNA-seq analysis with Pre-analysis and Core-analysis.

Raw reads

The raw reads need quality control process which includes the analysis of sequence quality, GC content, the presence of adaptors, overrepresented k-mers and duplicated reads due to detection of sequencing errors, PCR artifacts or contaminations. There are several quality control tools such as FastQC, NGSQC (Andrews 2010, Dai et al. 2010). FastQC is a popular tool to perform these analyses on Illumina reads, whereas NGSQC can be applied to any platform. Commonly, if read quality becomes too low, bases should be removed to improve mapping rate. Software tools such as the FASTX-Toolkit and Trimmomatic, which can be used to eliminate low-quality reads and poor-quality bases, and trim adaptor sequences (Gordon et al. 2010, Bolger et al. 2014).

Aligning reads to a reference genome

The computational analysis of an RNA-seq experiment begins in early: a set of FASTQ files that contain the nucleotide sequence of each read and a quality score at each position are first allowed. These reads must first be aligned to a reference genome or transcriptome. It is important to know if the sequencing experiment was single-end or paired-end, as the alignment software will require the user to specify both FASTQ files for a paired-end experiment. The output of this alignment step is commonly stored in a file format called SAM/BAM. The read mapping or alignment that is to find the

unique location of reads identical to the reference genome or transcriptome. For RNA-seq reads mapping, software tools such as BWA, Bowtie and Bowtie2 are commonly used earlier (Langmead et al. 2009, Li et al. 2010, Langmead et al. 2012), and the fast splice junction mapper TopHat and TopHat2 are widely used (Trapnell et al. 2009, Kim et al. 2013). TopHat2 aligns RNA-Seq reads to mammalian-sized genomes using the ultra-high-throughput short read aligner Bowtie2, and then analyzes the mapping results to identify splice junctions between exons. In addition, the others software tools such as ELAND, STAR, GSNAP, Rsubread, HISAT/HISAT2 and etc (Bentley et al. 2008, Nookaew et al. 2012, Dobin et al. 2013, Liao et al. 2013, Kim et al. 2015). STAR is a ultrafast universal RNA-seq aligner and can align reads in a continuous streaming mode which makes it compatible with novel sequencing technologies such as the one recently announced by Oxford Nanopore Technologies. Rsubread is an R/Bioc package that implements an extremely fast aligner for RNA-Seq data. It is currently only available for OS X and Linux, but not for Windows. HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads such as DNA and RNA to a population of human genomes, as well as to a single reference genome. While some studies compared performances among the align software tools, they concluded that it is impossible to determine specific align software tools is better in all conditions due to mapping rate is highly affected by many factors (i.e genome structure) (Grant et al. 2011). Thus, employing and comparing several align software tools, would be recommended in most RNA-seq studies. I summarized the performances such as sensitivity, precision, runtime and memory usage among aligners with reference to other

papers (Table 1.1) (Grant et al. 2011, Lindner et al. 2012, Dobin et al. 2013, Kim et al. 2013, Lu 2013, Fonseca et al. 2014, Kim et al. 2015, Medina et al. 2016).

Table 1.1. Comparison of performances such as sensitivity, precision, runtime and memory usage among aligners

Tools	Sensitivity (Rank)	Precision (Rank)	Runtime	Memory usage (GB)
Bowtie	10	7	Hours	NA
Bowtie2	11	10	Hours	NA
GMAP	13	11	Hours	NA
GNUMAP	12	9	Hours	NA
Tophat	9	8	Hours	2
Tophat2	8	6	Hours	4.3
HISAT	4	2	Minutes	4.3
STAR	6	5	Minutes	28
MapSplice2	2	4	Hours	3.3
HPG Aligner	1	1	Minutes	9
GSNAP	5	3	Hours	20.2
Olego	7	1	Hours	3.7
RUM	3	4	Hours	26.9

Summarizing mapped reads

After aligning reads to a reference genome, the next step is to summarize reads over already known exons, transcripts or genes. To quantify gene expression, the aligned reads data need to be translated into expression measurement by 1) counting the number of reads overlapping exons in a gene, 2) counting the number of the reads that include read along the whole length of the gene (incorporate introns as well), and 3) counting the number from de novo assembly of RNA-seq reads (Trapnell et al. 2009). When the first option is chosen, the reads information of mapped reads that locate outside already known exons is lost. When the second option is chosen, overlapping transcripts can be captured because there is sharing location (intron) by different genes. In the third option, to determine isoforms present within a sample, it requires deep sequencing. For RNA-seq reads counting, software tools such as summarizeOverlaps, featureCounts, tximport and htseq-count are commonly used (Anders et al. 2010, Lawrence et al. 2013, Liao et al. 2014, Soneson et al. 2015). In addition, for software tools that summarize the reads, htseq-count python module is used when RNA-seq reads are mapped to annotated coding regions, and Cufflinks and MiSO are used to estimate the proportion of reads that are assigned to splice variants and to identify isoforms (Katz et al. 2010, Trapnell et al. 2010). Thus, the aligned read count for each gene can be possibly changed by the choice of summarization, which should be selected based on the purpose. I summarized the performances such as sensitivity, precision, runtime and memory usage among transcript assemblers with reference to other papers (Table 1.2) (Kanitz et al. 2015, Pertea et al. 2015).

Table 1.2. Comparison of performances such as sensitivity, precision, runtime and memory usage among transcript assemblers

Tools	Sensitivity (Rank)	Precision (Rank)	Runtime	Memory usage (GB)
StringTie +SR	1	1	Minutes	6.225
StringTie	2	2	Minutes	6.125
Cufflinks	3	3	Hours	11.65
Traph	5	4	Days	NA
Scripture	3	5	Days	20.075
IsoLasso	4	5	Hours	13.85

Normalization

RNA-seq have two bias. first, Within sample, longer transcripts have higher read counts even though when their actual expression level is the same (Van Verk et al. 2013). Second bias, between samples, differences in sequencing depth makes individual sample bias. Therefore, RNA-seq read count data requires normalization within and between samples after summarization, in order to make an accurate comparison of expression levels. For normalization of RNA-seq read count data, software tools such as Total Counts (TC); Upper Quartile (UQ), Median (Med) , Trimmed Mean of M-values (TMM), Relative Log Expression (RLE) normalization using the DESeq package (DESeq), Quantile normalization (Q), Reads (single-read) or fragments (paired-end reads) per kilobase of exon model per million mapped reads (RPKM/ FPKM), and Remove Unwanted Variation (RUV) are commonly used (Bolstad et al. 2003, Smyth 2005, Mortazavi et al. 2008, Anders et al. 2010, Bullard et al. 2010, Robinson et al. 2010, Dillies et al. 2013, Risso et al. 2014). Among these normalization methods, RPKM/FPKM, TMM and RLE are popular. RPKM/FPKM is the most widely used method. However these normalization methods have some problems. One of the problem arises due to small number of highly expressed genes and genes length, and subsequently those genes are more likely to be detected as a differentially expressed genes, thus there is still bias for expression estimation in those genes (Bullard et al. 2010, Van Verk et al. 2013). TMM normalization is the EdgeR package's default normalization method, assumes that most genes are not differentially expressed and based on the negative binomial distribution (Robinson et al. 2010). It calculates a normalization factor for each gene, though this correction factor is applied to

library size (i.e. sequencing depth). Moreover, to compute the TMM factor, one lane is considered a reference sample and the others test samples, with TMM being the weighted mean of log ratios between test and reference. Then excluding the most expressed genes and the genes with the largest log ratios. For DESeq default normalization method, the RLE (Relative Log Expression) normalization method was used in EdgeR as it is equivalent. Starting with the hypothesis that most genes are not DE, scaling factors are calculated for each lane as median of the ratio, for each gene, of its read count of its geometric mean across all lanes. This way, non-differentially expressed genes will have similar read counts across samples, with a ratio of 1, and the median of the ratio for a given lane serves as a correction factor to apply to all read counts. The DESeq method and TMM outperformed Med, UQ, TC, Q, and RPKM normalization methods (Dillies et al. 2013).

Differential expression

The aim of the differential expression analysis in RNA-seq is to find genes that have significantly changed in expression across experimental conditions. To achieve this, expression in two or more samples needs to be calculated and the statistical significance of each observed change in expression between them needs to be tested. In RNA-seq, count data do not follow normal distribution. Therefore, a proper statistical model is used for normalization of count data. Differential expression analysis of RNA-seq data in the early days used, the Poisson distribution for normalization of RNA-seq count data (Robinson et al. 2010). However, this model cannot estimate the biological

variation well and is not proper for RNA-seq data (Langmead et al. 2010). To make flexible model estimate the biological variation, tests using the negative binomial distribution was developed (Robinson et al. 2007) , and another similar approaches such as the common dispersion model (Robinson et al. 2008), the empirical cumulative distribution functions method (Anders et al. 2010), the empirical Bayesian approach (Hardcastle et al. 2010), the two-parameter generalized Poisson model (Srivastava et al. 2010) and etc were introduced. Today, there are many software tools such as edgeR (Robinson et al. 2010), DESeq (Anders 2010), and Cuffdiff 2 within Cufflinks (Trapnell et al. 2013) that use negative binomial model. They are widely used and showed good performance (Kvam et al. 2012). EdgeR detects differential expression using empirical Bayes estimation and exact tests (i.e. Fisher's exact test) and based on a negative binomial model. The package has been developed to enable analysis of experiments with small numbers of replicates and without replicates. As default, the TMM normalization procedure is carried out to account for the different sequencing depths between the samples, whereas the Benjamini–Hochberg procedure is used to control the FDR (Benjamini et al. 1995). DESeq is similar to edgeR. First, it uses negative binomial model that more general then edgeR. Second, a scaling factor normalization procedure is carried out to account for the varying sequencing depths of the different samples and uses Benjamini–Hochberg procedure. Third, it can be possible to enable analysis of experiments with small numbers of replicates and without replicates, but it is technically possible, although not recommended, to work with experiments without any biological replicates. Cuffdiff 2 is part of the extensive Cufflinks package developed for the identification of differentially

expressed genes and transcripts and revealing differential splicing and promoter-preference changes. It estimates expression at transcript-level resolution and controls for variability and read mapping obscurity by a beta negative binomial model for fragment counts. As default, Cuffdiff 2 uses a similar scaling factor procedure as DESeq. The Cuffdiff 2 method specifically addresses the ambiguity in counts due to obscure reads that result in false differential expression genes especially with several similar isoforms. In addition, the others software tools such as NOIseq (Tarazona et al. 2011), SAMseq (Li et al. 2013), Limma (Smyth 2005), EBSeq (Leng et al. 2013), baySeq (Hardcastle et al. 2010), voom (Law et al. 2014), NBPSeq (Di et al. 2011), TSPM (Auer et al. 2011) and ShrinkSeq (Van De Wiel et al. 2012). BaySeq is based on estimating posterior likelihoods of differential expression via empirical Bayesian methods, assuming negative binomially distributed data. The method produces posterior probabilities rather than significance values and reports a Bayesian FDR estimate. Limma is based on linear modeling. It was originally designed for analyzing microarray data but has recently been extended to RNA-seq data. The current recommendation according to the limma user guide is to use TMM normalization of the edgeR package and the so called ‘voom’-conversion which essentially transforms the normalized counts to logarithmic (base 2) scale. This method estimates the mean–variance relationship of the normalized counts to determine a weight to each observation prior to linear modeling. By default, the Benjamini–Hochberg procedure is used to estimate the FDR. Therefore, I summarized the features and performances among differentially expression methods with reference to other papers (Table 1.3) (De Paepe 2015, Frazee et al. 2015,

Leon-Novelo et al. 2015, Seyednasrollah et al. 2015, Tarazona et al. 2015, Gim et al. 2016, Pimentel et al. 2016).

Table 1.3. Comparison selected differentially expression methods

Normalization, Quantitative analysis and Differential Expression tools	EdgeR	DESeq	DESeq2	limmavoom	Balllgown	cuffdiff2	EBSeq	baySeq	PoissonSeq	NOIseq	SAMseq
Quantification measure	Count-based	Count-based	Count-based	Count-based, linear model	Linear model	Count-based	Count-based, Linear model	Count-based	Count-based	Count-based	Count-based
Normalization	TMM/Upper quartile/ RLE (DESeq-like)/None (all scaling factors are set to be one)	Median-of-ratio	Median-of-ratio	TMM	FPKM	Geometric (DESeq-like) /quartile/classic-fpkm	Median Normalization	Scaling factors (quantile/TMM/total)	Total count of least differental genes (assessed by GOF)	RPKM/TMM/Upper quartile	Poisson Sampling
Read count distribution assumption	Negative binomial distribution	Negative binomial distribution, Poisson distribution (no or few replicates)	Negative binomial distribution	Negative binomial distribution	Beta negative binomial distribution	Negative binomial distribution	Negative binomial distribution	Negative binomial distribution	Negative binomial distribution	Nonparametric method, empirical distribution (no or few replicates)	Nonparametric method
Differential expression test	Exact test	Exact test	Exact test	Empirical Bayes method	Parametric F-test comparing nested linear models	t-test	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods	Score statistic on the basis of the a Poisson log limear model	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null	Wilcoxon rank statistic and a resampling strategy
Support for multi-factored experiments	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
True positive rate	High	Low	Low/Medium	Low/Medium	Medium/High	Low	Independent of sample size	Low	High	Not clear	Low(small sample sizes)/ High(large

											enough sample sizes)
Support differential express detection without replicated samples	Yes	Yes	No	No	No	Yes	No	No	Yes	Yes	No
Detection of differential isoforms	No	No	No	No	Yes	Yes	Yes	No	No	No	No
Runtime for experiments	Minutes	Minutes	Minutes	Minutes	Seconds (standard laptop)	Hours	Hours	Hours	Seconds	highly dependent on sample size	highly dependent on sample size

De novo assembly

De novo assembly can be used when official reference genome is not available, often in the non-model species. In this situation, de novo assembly based RNA-seq analysis is the best alternative solution for quantification of gene expression level. By assembling among the RNA-seq read, not genome reference, but transcriptome reference, can be constructed. Since, RNA-seq analysis can be implemented similarly as the reference genome based approach. In the transcriptome assembly, the results are highly affected by gene expression levels of the targeted samples. For example, highly expressed transcriptome can be well constructed based on its high depth coverage. On the contrary, transcriptome with the low expression level (low depth coverage region) cannot be as well assembled as the highly expressed region, which is one of the limitations in assembly based approach. To solve this problem, several computational methods were suggested in the RNA-seq analysis field. Currently, the most popular software packages for short-read RNA-seq de novo assembly include Oases (Zerbino et al. 2008, Schulz et al. 2012), Trinity (Grabherr et al. 2011, Haas et al. 2013), trans-ABySS (Simpson et al. 2009, Robertson et al. 2010) and SOAPdenovo-Trans (Xie et al. 2014). All four packages are based on constructing, simplifying, and resolving de Bruijn graphs to extract likely transcripts (Compeau et al. 2011). Two of these, Oases and trans-ABySS start with constructing de Bruijn graphs directly from sequencing reads, remove potential errors, and then resolve each de Bruijn graph to extract transcripts for each connected component (i.e. cluster, or "locus") in the graph. Both packages use a range of k-mer sizes to accommodate variation in read coverages among genes. Trinity, on the other

hand, uses a single k-mer with size fixed at 25 bp. Trinity first carries out a greedy extension step starting from the most abundant k-mer to build linear contigs, groups overlapping contigs into connected components, and constructs a de Bruijn graph for each component. Sequencing reads are then mapped to the graphs, the graphs are simplified, errors are removed (which may break a component into subcomponents), and finally predicted isoforms are extracted for each component or subcomponent. All four also use the information from mate pairs to assemble contigs into scaffolds when paired end reads are available. Each "locus" from the Oases output (roughly equivalent to component/subcomponent in Trinity) consists of one or more "transcripts" (or "isoforms" in Trinity). Biologically a locus or a component/subcomponent can each contain one gene or several paralogs, and a single gene can have fragments distributed among multiple loci or components/subcomponents. Trans-ABySS does not explicitly output sequences in hierarchical groups. SOAPdenovo-Trans is available only as precompiled executables without formal publication or source code. All the published de novo transcriptome assemblers are optimized for building references for comparing gene expression levels, identifying splice variants, and determining gene fusion events. Hybrid approach between reference genome and de novo based approaches can be implemented using cufflinks that provides comprehensive analyzing pipeline including aligner, assembler, functional annotation, splice sites discovery and statistical analysis. Assembled results in each of the software tools differs, which is already reported (Jain et al. 2013). Moreover, Trinity and Tophat–Cufflinks combination, were better in terms of recovery rate and length of assemble

contigs than other combinations. Notably, when there is reference genome, genome-guided Trinity and Cufflinks is best combination for high accuracy of identification of transcriptome contigs, more accurate transcript assemblers have been developed with Bridger (Chang et al. 2015)

1.3 Evolution of Domestic Animal

1.3.1 Definitions and history

In a brief definition, domestic animal including the horse or cat, that has been tamed and kept by humans used in various ways, such as food source, or pet. As a result of selective breeding, domestic animals have become notably different from their wild ancestors. The domestication of animals is the scientific theory of the mutual relationship between animals and humans who control their safety and reproduction (Zeder 2015). There is not only a genetic difference between domestic and wild populations, but also differences between the domestication traits and enhancement traits. The domestication traits are believed to have been essential at the early stages of domestication by researchers, and the enhancement traits have appeared since the isolation between wild and domestic populations (Olsen et al. 2013, Doust et al. 2014, Larson et al. 2014). From about 20,000 years ago with most recent ice age, large mammals such as bison inhabited the sub-arctic tundra of Europe and Asia. They were preyed upon by two groups of hunters such as humans and wolves. About 12,000 years ago, the earliest known evidence of a domesticated dog is a jawbone found in a cave in Iraq. Over the past 11,500 years, the domestication of plants and animals has significantly transformed Earth's biosphere, which caused the changing human population size and evolution. Animal domestication has taken place over timescales accessible through archaeological evidence and been driven by selection pressures created by both unintentional and deliberate human actions. In 1868 years,

Darwin was the first to know that domestic animals possess a wide variety of similar morphological traits despite the lack of close evolutionary relationships between their wild ancestors and the difference between conscious selective breeding in which humans directly select for desirable traits and unconscious selection where traits evolve as a by-product of natural selection or from selection on other traits (Darwin 1868, Jared 1997, Larson et al. 2014) In 1907 years, Francis Galton suggested that dogs were domesticated following the capture and nurturing of wolf puppies in human camps. He based this conclusion on ethnographic research that suggested domesticated dog is called pet, was not unusual among hunter and gatherer groups across the globe. Even if some scientists felt this observation did not constitute a sufficient explanatory mechanism (Serpell 1989). In the 1950s, Dmitry Belyaev had identified that the appearance of the domestication syndrome and how it could have resulted without both human intentionality and selection pressures focused upon individual traits by using silver foxes (Trut 1999, Trut et al. 2009). In the 2010s, Vigne proposed a multistage model that was characterized by a gradually enhancing relationship between humans and animals. In this view, animal domestication proceeded along a continuum from anthropophilia to commensalism, to control in the wild, to control of captive animals, to extensive and intensive breeding, and finally to pets (Vigne 2011). Although Zeder also knew the idea of the staged model approach, her described three separate pathways that animals followed in the domestication by human: a commensal pathway, a prey pathway, and a directed pathway (Zeder 2012).

1.3.2 Horse

Humans acquire their most important single ally from the animal kingdom when they domesticate the horse, in about 3000 BC. Wild horses of various kinds have spread throughout most of the world by the time human history begins. Their bones feature among the remains of early human meals, and they appear in cave paintings with other animals of the chase. Some of their earliest fossil remains have been found in America, but after arriving across the Bering Land Bridge they become extinct in that continent. They are reintroduced by European colonists in the 16th century. A natural habitat of the wild horse is the steppes of central Asia. Here, with its ability to move fast and far, it can gallop out of harm's way and make the most of scarce grazing. And here, human's first capture, tame and breed the horse is predicted to have taken place approximately 5000 years ago. The original purpose, as with cattle, is to acquire a reliable source of meat and subsequently milk. But then, in a crucial development, tribesmen discover that they have at their disposal a means of transport. With a horse beneath him, man's ability to move is improved out of all recognition. The next comparable moment in the story of human speed does not arrive for another 5000 years - with steam trains. The first domesticated horses are of a size which I would describe as ponies. Horses of this kind were still living in the wild in Mongolia until quite recent times. Discovered there in the 1870s, and named Przewalski's horse, they survive now only in zoos. Thoroughbred, breed of horse developed in England for racing and jumping. The origin of the Thoroughbred may be traced back to records indicating that a stock of Arab and Barb horses was introduced into England as early as the 3rd century.

1.3.3 Pig

Pigs are associated with settled communities, which are domesticated slightly later, but probably not long after 7000 BC. The pig is probably first domesticated in China. The first reason for keeping pigs in the village, is to secure a regular supply of fresh meat. The hunter is dependent on the luck of the chase; if more animals are killed than can be immediately consumed, meals from the surplus will be increasingly unpleasant as the days go by. The herdsman, by contrast, has a living larder always to hand and a supply of dairy products as well. Pig also provide for almost every other need of neolithic man. While they are alive, they produce dung to manure the crops. When they are dead, leather for garments; bone for sharp points, of needles or arrows; fat for tallow candles; hooves for glue.

1.3.4 Chicken

The red jungle fowl, a member of the pheasant family, lives in the forests and bamboo jungles of India and south East Asia. The male makes an impressive crowing sound and is dignified by a comb on his head and wattles under his beak. Jungle fowl of this kind are captured and kept for their eggs and their flesh by about 2000 BC in Asia. It is thought that all domestic poultry in the world today are descended from this one species. At much the same period, in Egypt, pigeons are first persuaded to live and breed in the proximity of humans - again as a reliable source of protein. But some 3000 years later it is

discovered that they have an extra and unusual talent. Some of them can be trained to fly home.

This chapter was published in *Molecular biology Report*
as a partial fulfillment of Woncheoul Park's Ph.D program.

Chapter 2. Comparative transcriptomic analysis to identify differentially expressed genes in fat tissue of adult Berkshire and Jeju Native Pig using RNA-seq

2.1 Abstract

Jeju native pigs (JNP) have been adapted to an exotic natural environmental niche. They have been known to be resistant disease and have a good meat quality because of higher tenderness, juiciness, redness and brightness than those of Western breeds. In order to understand the molecular mechanisms of JNP specific phenotype, here I conducted comparative transcriptomics study using RNA-seq technology. I compared transcriptome between JNP and Berkshire in three different tissues (fat, liver and muscle). I identified differential expressed genes (DEGs) of each tissue between the two breeds. Among the DEGs, I found that 26 genes were related to meat quality and body growth. Among those genes, *MPZ*, *AADAT*, *IGFN1* and *MYBPH* were up-regulated in JNP. Therefore, I suggest that JNP has different gene expression profile which related to meat quality and body growth compared to Berkshire.

2.2 Introduction

The pig (*Sus scrofa domesticus*) was domesticated from the wild boar approximately 9,000 years ago (Kijas et al. 2001, Larson et al. 2005). It has become an important animal as one of the major animal protein sources for humans. An exotic natural environment of Jeju island of Korea differs from Korean peninsula so that it contains several Jeju native livestock resources, such as horses, chicken, cattle and pigs. Among these, JNP have been adapted to the unique environmental niche. Historically it is estimated that JNP was introduced in 12th century from main China. They have preferred taste for Korean compared with other commercial breed because higher tenderness, juiciness, redness and brightness than those of Landrace and Western breeds (Jin et al. 2001, Cho et al. 2011). In addition, it shows strong resistance to disease (Kim et al. 2009). However JNP show lower feed efficiency and smaller litter size (5~8 litter size).

Fat and variation in fatty acid contents affects meat quality. Due to their different melting points, variation in fatty acid contents have an important effect on the firmness or softness of the fat in meat and thereby affect meat quality, especially the subcutaneous and intermuscular (carcass fats), but also the intramuscular (marbling) fat and content (IMF) (Urban et al. 2002, Wood et al. 2004). Moreover, lipogenesis in pig occurs in both the liver and adipose tissue (Azain 2004). Liver is the major site of fatty acid synthesis (Corino et al. 2002). and liver lipogenesis is related to body growth in the pig (Mourot et al. 1995). Therefore, liver tissue is also a major factor in meat quality. Another

major factor in meat quality and body growth is muscle, specifically, muscle composition, areas of fiber and the capillary density of specific muscles are important factors influencing many peri-mortal and post-mortal biochemical processes and thereby meat quality (Klont et al. 1998, Malek et al. 2001)

In prior studies, genes and markers of differences between pig breeds have been identified using microarray, qRT-PCR and microsatellite analyses (Kim et al. 2005, Park et al. 2007, Kim et al. 2008, Moon et al. 2009). Recently, methods for transcriptome profiling using deep-sequencing and short-read technologies (RNA-seq) were developed. RNA-seq technology enables more comprehensive investigation of the transcriptome than microarrays and is becoming more popular for gene expression studies (Mortazavi et al. 2008, Nagalakshmi et al. 2008) . Transcriptome analysis using RNA-seq has been applied recently in other pig breeds, but not JNP (Chen et al. 2011, Petkov et al. 2011, Rustemeyer et al. 2011, Jung et al. 2012, Looft 2013, Prather 2013, Samborski et al. 2013). Indeed, comparisons of the growth and meat quality of other pig breeds are scarce.

Statistical analysis is critical in transcriptome studies using RNA-seq; specifically, the normalization of quantitative measurements of expression (Wilhelm et al. 2009, Bullard et al. 2010, Li et al. 2010, Robinson et al. 2010, Hong et al. 2012), as well as detection of differentially expressed genes (Hardcastle et al. 2010, Robinson et al. 2010, Wang et al. 2010, Tarazona et al. 2011, Trapnell et al. 2012).

Herein, I performed a transcriptome analysis using RNA-seq on three tissue types related to meat quality and growth: fat, liver and muscle harvested from adult JNP and adult Berkshire. I detected DEGs and identified important

genes related to meat quality and growth in JNP. To our knowledge, this study is the first statistical analysis to detect DEGs from RNA-seq data from a small sample with no replicates. Moreover, I identified many important candidate genes in JNP related to meat quality and growth by DEG profiling.

2.3 Materials and Methods

2.3.1 Animals and sample preparation

Animals under study were from JNP and Berkshire breeds. They were housed in similar environmental and nutritional conditions. Animals were slaughtered according to the standard protocols of Jeju National University and the fat samples were collected after slaughter. Samples were stored immediately in dry ice and later were stored at -80℃ until used for RNA extraction. The research proposal and the relevant experimental procedures were approved by the institutional review board of the Department of Animal Biotechnology, Jeju National University.

2.3.2 Extraction and analysis of quality of RNA

RNA was isolated from 100 mg of the fragmented frozen fat tissue samples from adult JNP and Berkshire pigs. TRIzolTM (Invitrogen, USA) reagent was used for the isolation of RNA. Tissue samples were homogenized in 1.5 ml of TRIzol reagent and chloroform, which were subsequently precipitated by using isopropanol (Junsei Chemical Co. Ltd., Japan). Isolated RNA samples were stored at -80℃. To purify RNA from genomic DNA contamination, 25 lg of RNA from each sample was treated with the RNase-free DNase set (QIAGEN, Hilden, Germany) and it was purified with the RNeasy mini kit according to the user guidelines (QIAGEN, Hilden, Germany). A Bioanalyzer 2100 with RNA 6000 Nano Labchips was used to assess the quality and

quantity of RNA by automated capillary gel electrophoresis by following user guidelines (Agilent Technologies Ireland, Dublin, Ireland). 28S/18S ratios for the RNA samples ranged from

2.3.3 Analysis of RNA-seq reads and identification of DEGs

The quality of RNA-seq reads from the three JNP and Berkshire pig tissues was checked using FastQC (Supplementary Figure S1). Reads that passed the quality control were mapped to the Sus scrofa genome (Sscrofa10.2) from UCSC using Tophat2 (v2.0.2) and reads were counted using HTseq (v0.5.3p3). I used the corrplot (Friendly 2002) R package to analyze the correlation between adult JNP and adult Berkshire. I used the DEGseq (Wang et al. 2010) R package to identify DEGs between JNP and Berkshire pig from the raw count dataset. Using this statistical model, Fisher's exact test was used to identify DEGs (Bloom et al. 2009). Because DEGseq enables detection of DEGs between two samples, I compared the two breeds in pairs (i.e., between JNP fat and Berkshire fat, JNP muscle and Berkshire muscle, and JNP liver and Berkshire liver). Significant DEGs were selected at FDR <0.01.

2.3.4 Functional annotation of DEGs

The bovine Ensembl gene IDs were converted to official gene symbols by cross-matching to human Ensembl gene IDs and official gene symbols. The official gene symbols of human homologs of bovine genes were used for

functional clustering and enrichment analyses using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis Jr et al. 2003). The representations of functional groups in fat, liver and muscle tissue between JNP and Berkshire relative to the whole genome were investigated using the Expression Analysis Systematic Explorer (EASE) tool (Hosack et al. 2003) within DAVID, which is a modified Fisher's exact test used to measure the enrichment of gene ontology (GO) terms (Alterovitz et al. 2010). To identify enriched GO terms, functionally clustered genes were filtered according to EASE values <0.1, and selected.

2.4 Result

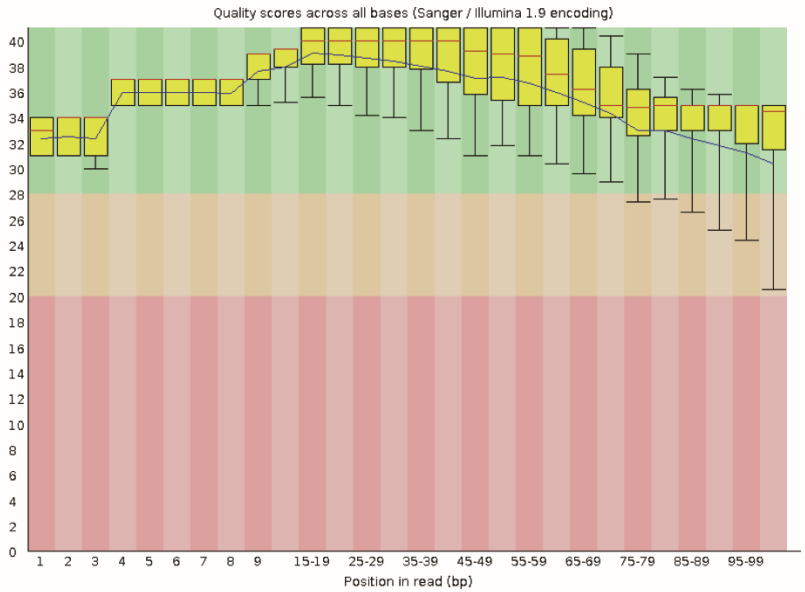
2.4.1 Quality of RNA-sequence reads in three tissues between JNP and Berkshire

I acquired RNA-seq reads in three tissues (fat, liver, Muscle) from two pig breeds such as Berkshire and Korean native [GSE45204]. The numbers of total sequence reads and mapping rates for each sample are shown in Table 2.1. Few sequence reads (<5.0E-05%) did not pass the quality filtering. The average numbers of sequence reads in Berkshire and JNP pigs were 37, 40 and 33M in fat, liver and muscle, respectively. Among the sequence reads that passed the quality control, on average, 90.7% reads in fat, 97.6% in liver, and 90.8% in muscle were mapped successfully to the pig genome (Sscrofa10.2) using TopHat (v2.0.2) (Table 2.1).

Table 2.1. RNA-seq reads and mapping rate of different tissue from KNP and Berkshire in pig breeds

Pig breeds	No. of	Tissue		
		Fat	Liver	Muscle
KNP	total reads	41413088	45255914	35563538
	reads after QC	41411459	45254078	35562180
	accepted Hit	37879239 (91.5%)	44164596 (97.6%)	31807200(89.4%)
Berkshire	total reads	40310004	37618038	37310958
	reads after QC	40308452	37616584	37309517
	accepted Hit	36254763(89.9%)	36731445(97.6)	34331014 (92.0%)

a)Per base sequence quality



b)Per base N content

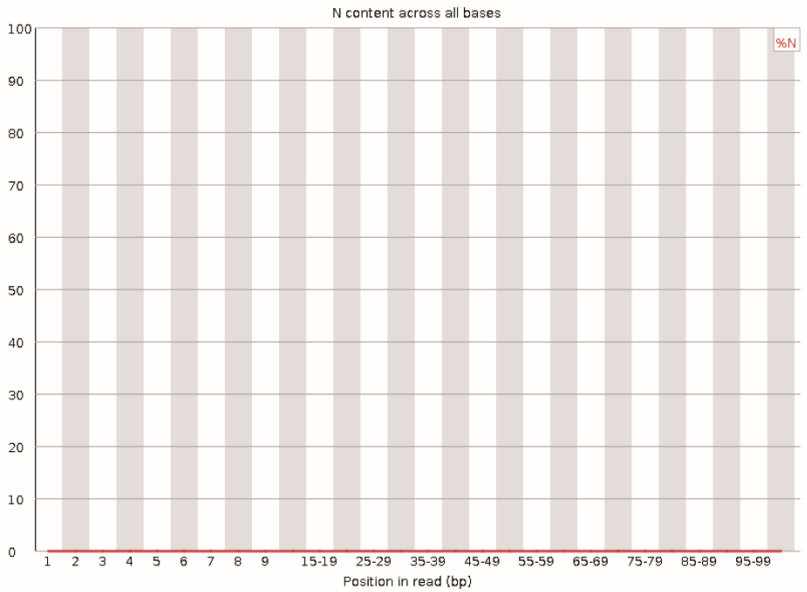


Figure 2.1. Data quality control using FastQC

2.4.2 Identification of differentially expressed genes (DEGs)

I determined correlations between different tissues in JNP and Berkshire. Strong correlations between JNP and Berkshire were identified in fat, liver and muscle (Figure 2.2). I identified DEGs using the expression profiles of genes in fat, liver, and muscle tissue of JNP and Berkshire pigs. I identified 153 (87 up-regulated, 66 down-regulated), 169 (90 up-regulated, 79 down-regulated) and 39 (17 up-regulated, 22 down-regulated) DEGs in fat, liver and muscle, respectively, differentially expressed between JNP and Berkshire (FDR <0.01). Of these DEGs, 96 (50 up-regulated, 46 down-regulated), 99 (45 up-regulated, 54 down-regulated) and 22 (12 up-regulated, 10 down-regulated) from fat, liver and muscle, respectively, have been annotated (Table 2.2 and Figure 2.3). Open source online visualization of these data is available (http://biopopdb.snu.ac.kr/PIG_DEG/).

Table 2.2. Summary of DEG identified from three different tissues between JNP and Berkshire (FDR<0.01).

		Tissue		
		Fat	Liver	Muscle
No. of DEGs	Up-regulated	87	90	17
	Down-regulated	66	79	22
	Total	153	169	39
No. of annotated gene	Up-regulated	50	45	10
	Down-regulated	46	54	12
	Total	96	99	22

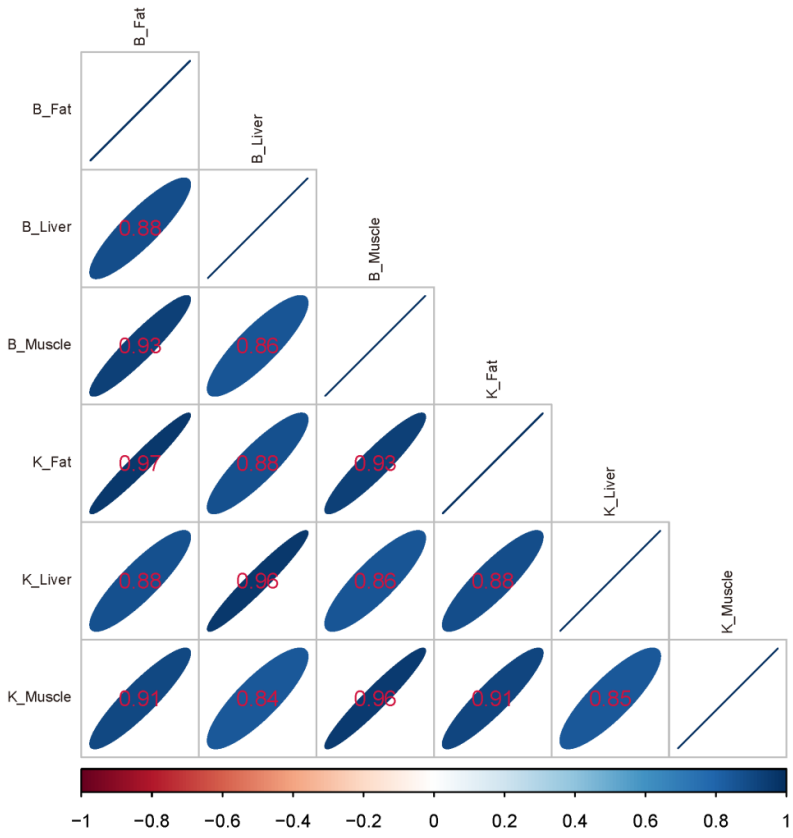


Figure 2.2. Correlation plot between KNP and Berkshire

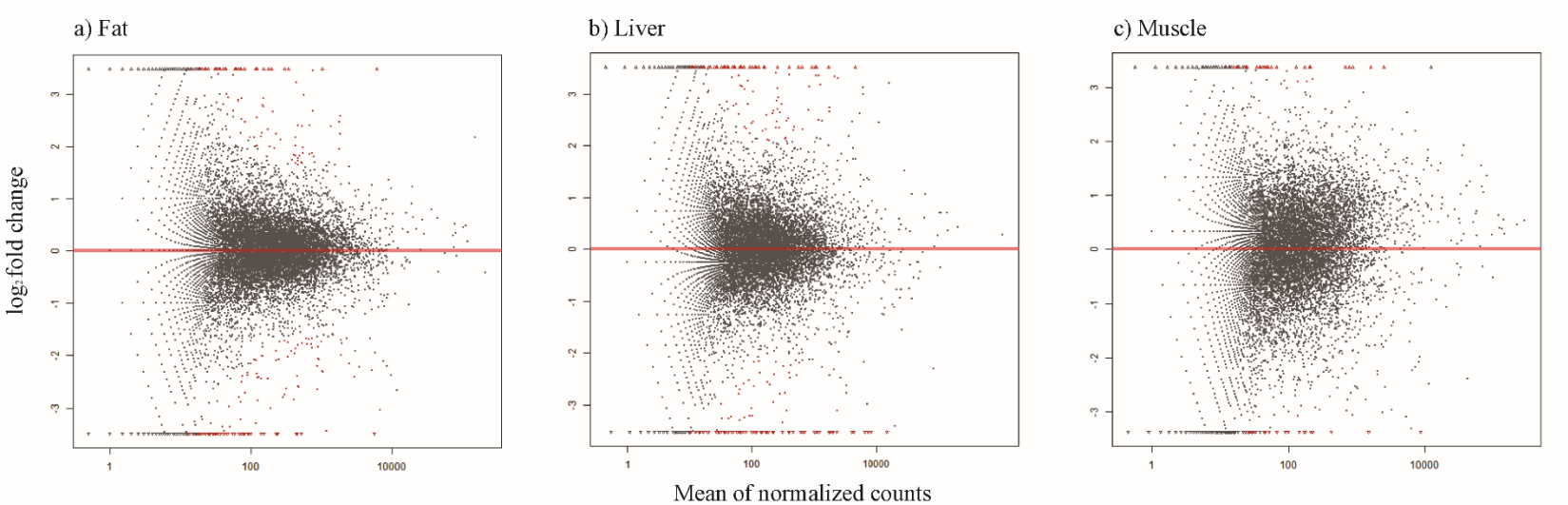


Figure 2.3. MA plot between KNP and Berkshire

2.4.3 DEGs involved in the meat quality and body growth

Of the DEGs, 26 were related to meat quality and body growth (Table 2.3). *COL11A1*, *COL11A2* in fat and *COL2A1* in muscle, which are related to cartilage growth (Lefebvre et al. 2005), were up-regulated. *KERA*, *TNMD* and *PCK1*, related to muscle growth (Srikanchai et al. 2010, Karasik et al. 2012, Zhai et al. 2012), were up-regulated in fat. *PDK4*, associated with meat quality of skeletal muscle (Lan et al. 2009), was up-regulated in fat and muscle. *TBX15*, related to dorsal-ventral distribution of subcutaneous adipose tissue (Komolka et al. 2012), was up-regulated in fat. *ACAA2*, *APOA4*, *CYP2E1* and *HMGCS2*, related to hepatic lipid, HDL and growth hormone (Zhou et al. 2006, Daniels 2009, Uddin et al. 2011, Graugnard et al. 2012, Mörlein et al. 2012), were up-regulated in liver. *MT3*, related to growth inhibition (Cao et al. 2004), was up-regulated in liver. *LEPR*, related to fat content and body weight (Muñoz et al. 2011, Tyra et al. 2011), was up-regulated in liver. *PNPLA3*, related to fat deposition (Chen et al. 2011), was down-regulated in liver. *POSTN*, related to muscle growth (Bílek et al. 2008), was up-regulated in muscle. *HOXC6*, *HOXC8* and *HOXD8*, of the gene family HOX known to be related to bone and mammary gland development (Visvader et al. 2003, Lindholm-Perry et al. 2010), were down-regulated in muscle. *IGFN1* and *MYBPH*, related to myosin structure (Morzel et al. 2008, Zhang 2009), were down-regulated in muscle.

Table 2.3. Identified DEGs related to meat quality and body growth in three tissues

Breed	Tissue	Gene	Related fuction	Refernece
JNP	Fat	COL11A1	BG	V Lefebvre and P Smits, 2005
		COL11A2	BG	V Lefebvre and P Smits, 2005
		KERA	MQ	T Srikanchai et al, 2010
		PCK1	BG	W. Zhai et al, 2012
		PDK4	MQ	Jing Lan et al, 2009
		TBX15	BG	Katrin Komolka et al, 2012
		TNMD	BG	David Karasik and Miri Cohen-Zinder, 2012
	Liver	ACAA2	MQ	Muhammad J Uddin et al, 2011
		APOA4	BG, MQ	H Zhou et al, 2006 and TF Daniels et al, 2009
		CYP2E1	MQ	D Mörlein et al, 2012
		HMGCS2	BG	D.E. Graugnard et al, 2012
		HSD17B2	BG, MQ	Xiaoping Li et al, 2010 and Shen Zhongyi et al, 2007
		LEPR	BG, MQ	G. Muñoz et al, 2011 and M. Tyra et al, 2007
		MT3	BG	H. Cao et al, 2004 and A Flores-Morales et al, 2008
		SLC13A5	MQ	W Luo et al, 2012
	Muscle	COL2A1	BG	V Lefebvre and P Smits, 2005
		PDK4	MQ	Jing Lan et al, 2009
		POSTN	BG, MQ	K Bílek et al, 2008
Berkshire	Fat	MPZ	BG	D. Wagenknecht et al, 2010
		S100A6	BG	DS Sisk , 2009
	Liver	AADAT	BG	CP Cabrera et al, 2011
		PNPLA3	BG, MQ	Zhilong Chen et al, 2011
		IGFN1	MQ	W Zhang et al, 2009
	Muscle	MYBPH	MQ	Martine Morzel et al, 2008
		HOXC6	BG	AK Lindholm-Perry et al, 2010 and JE Visvader et al, 2003
		HOXC8	BG	AK Lindholm-Perry et al, 2010 and JE Visvader et al, 2003
		HOXD8	BG	AK Lindholm-Perry et al, 2010 and JE Visvader et al, 2003

2.4.4 Gene ontology and functional annotation of DEGs

I summarized the biological process gene ontology of DEGs in three tissues of JNP and Berkshire pigs (Figure 2.4). I also summarized the biological process gene ontology of up-regulated DEGs in three tissues of JNP and Berkshire pigs (Figure 2.6). Embryonic morphogenesis and skeletal system development were the most significantly enriched groups in fat tissues of JNP and Berkshire pigs (p = 3.47E-04 and p = 4.48E-04, respectively) (Figure 2.4a). Cholesterol metabolic processes and sterol metabolic processes were the most significantly enriched groups in liver tissue (p = 3.90E-08 and p = 8.14E-08, respectively) (Figure 2.4b). Skeletal system development and skeletal system morphogenesis were the most significantly enriched groups in muscle tissue (p = 6.03-08 and p = 3.45E-04, respectively) (Figure 2.4c). In addition, I summarized the biological process gene ontology of specific up-regulated DEGs in three tissues of JNP and Berkshire pigs (Figure 2.5). Multicellular organismal processes and developmental processes were the most significantly enriched terms in fat tissues of JNP and Berkshire pigs (p = 1.17E-04 and p = 0.044, respectively). Metabolic processes was the most significantly enriched term in liver (p p = 0.001 and p = 0.03, respectively). Developmental processes was the most significantly enriched term in muscle from JNP (p = 0.02). Moreover, I summarized the cellular components and molecular function gene ontology of DEGs in three tissues of JNP and Berkshire pigs (Table 2.4). The cellular components gene ontology of specific DEGs were related to the extracellular region, endoplasmic reticulum, and proteinaceous extracellular matrix in fat, liver and muscle, respectively.

Further, the molecular function gene ontology of specific DEGs was related to cofactor binding in fat and liver tissues, respectively.

Table 2.4. GO terms of cellular components and molecular function of three tissues specific DEGs

(a) Up-regulated DEGs

Tissue		GO ID	Term	Count	PValue
Fat	Cellular Component	GO:0044421	extracellular region part	10	0.002870706
		GO:0005578	proteinaceous extracellular matrix	6	0.003410792
		GO:0031012	extracellular matrix	6	0.004696138
		GO:0005581	collagen	3	0.005586028
		GO:0005592	collagen type XI	2	0.006405233
		GO:0005576	extracellular region	14	0.008063527
		GO:0005583	fibrillar collagen	2	0.037835765
		GO:0044420	extracellular matrix part	3	0.054045365
	Molecular Function	GO:0048037	cofactor binding	5	0.002444996
		GO:0004499	flavin-containing monooxygenase activity	2	0.011502063
		GO:0050660	FAD binding	3	0.011934489
		GO:0046872	metal ion binding	16	0.029411119
		GO:0043169	cation binding	16	0.031961361
		GO:0046983	protein dimerization activity	5	0.034829846
		GO:0043167	ion binding	16	0.036367379

		GO:0050662	coenzyme binding	3	0.065182175
		GO:0008237	metallopeptidase activity	3	0.066447656
		GO:0050661	NADP or NADPH binding	2	0.08001191
		GO:0046982	protein heterodimerization activity	3	0.082939829
Liver	Cellular Component	GO:0044432	endoplasmic reticulum part	5	0.009415418
		GO:0005789	endoplasmic reticulum membrane	4	0.026839645
		GO:0042175	nuclear envelope-endoplasmic reticulum network	4	0.030844928
		GO:0005576	extracellular region	10	0.04429072
		GO:0031090	organelle membrane	7	0.045143295
		GO:0034364	high-density lipoprotein particle	2	0.058953681
		GO:0044421	extracellular region part	6	0.078955976
		GO:0005783	endoplasmic reticulum	6	0.078955976
		GO:0034358	plasma lipoprotein particle	2	0.081580914
		GO:0032994	protein-lipid complex	2	0.081580914
	Molecular Function	GO:0048037	cofactor binding	6	4.40E-04
		GO:0008483	transaminase activity	3	0.001355007
		GO:0016769	transferase activity, transferring nitrogenous groups	3	0.002765698
		GO:0005506	iron ion binding	5	0.008222674
		GO:0030170	pyridoxal phosphate binding	3	0.008748094

		GO:0070279	vitamin B6 binding	3	0.008748094
		GO:0009055	electron carrier activity	4	0.019754313
		GO:0050997	quaternary ammonium group binding	2	0.023331006
		GO:0019842	vitamin binding	3	0.045314303
		GO:0070330	aromatase activity	2	0.063510394
		GO:0043178	alcohol binding	2	0.063510394
		GO:0043498	cell surface binding	2	0.073301783
		GO:0016712	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	2	0.075734062
		GO:0050662	coenzyme binding	3	0.08111459
		GO:0050661	NADP or NADPH binding	2	0.09019814
Muscle	Cellular Component	GO:0005578	proteinaceous extracellular matrix	3	0.003165705
		GO:0031012	extracellular matrix	3	0.003766135
		GO:0044421	extracellular region part	3	0.020706272
		GO:0005576	extracellular region	3	0.082339447

(b) Down-regulated DEGs

Tissue		GO ID	Term	Count	PValue
Fat	Cellular Component	GO:0005576	extracellular region	14	0.002266435
	Molecular Function	GO:0030246	carbohydrate binding	7	6.00E-04
		GO:0005509	calcium ion binding	9	0.005249979
		GO:0001871	pattern binding	4	0.010945127
		GO:0030247	polysaccharide binding	4	0.010945127
		GO:0005529	sugar binding	4	0.020512922
		GO:0016918	retinal binding	2	0.035472287
		GO:0004806	triacylglycerol lipase activity	2	0.047021614
		GO:0005540	hyaluronic acid binding	2	0.055595071
		GO:0005501	retinoid binding	2	0.061268907
		GO:0005539	glycosaminoglycan binding	3	0.066025657
		GO:0019840	isoprenoid binding	2	0.066909525
		GO:0048306	calcium-dependent protein binding	2	0.091886278
Liver	Cellular Component	GO:0005792	microsome	4	0.030783085
		GO:0042598	vesicular fraction	4	0.033146771
		GO:0005829	cytosol	9	0.033584578
		GO:0005789	endoplasmic reticulum membrane	4	0.042344412
		GO:0042175	nuclear envelope-endoplasmic reticulum network	4	0.048421425

		GO:0044432	endoplasmic reticulum part	4	0.078292885
	Molecular Function	GO:0030554	adenyl nucleotide binding	14	0.001011992
		GO:0001883	purine nucleoside binding	14	0.001166743
		GO:0001882	nucleoside binding	14	0.00124411
		GO:0005524	ATP binding	13	0.001905459
		GO:0032559	adenyl ribonucleotide binding	13	0.00213891
		GO:0000166	nucleotide binding	16	0.003219853
		GO:0017076	purine nucleotide binding	14	0.005930387
		GO:0048037	cofactor binding	5	0.008348941
		GO:0032553	ribonucleotide binding	13	0.011290924
		GO:0032555	purine ribonucleotide binding	13	0.011290924
		GO:0050662	coenzyme binding	4	0.020549532
		GO:0016410	N-acyltransferase activity	3	0.029473373
		GO:0016878	acid-thiol ligase activity	2	0.062794229
		GO:0016877	ligase activity, forming carbon-sulfur bonds	2	0.083849328
		GO:0017137	Rab GTPase binding	2	0.09273093
		GO:0016769	transferase activity, transferring nitrogenous groups	2	0.09273093
Muscle	Molecular Function	GO:0043565	sequence-specific DNA binding	4	0.003090097
		GO:0003700	transcription factor activity	4	0.011746757

GO:0030528	transcription regulator activity		4	0.038454308

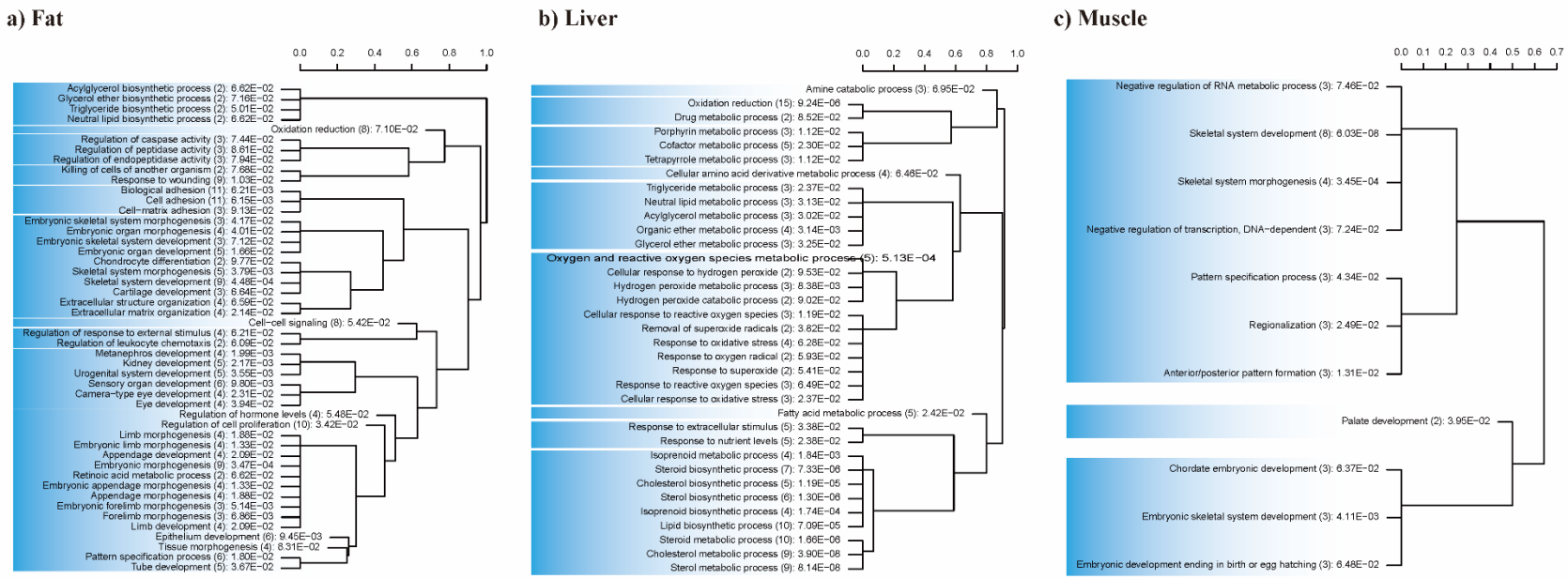


Figure 2.4. Enriched biological process GO terms of three tissues specific DEGs between JNP and Berkshire

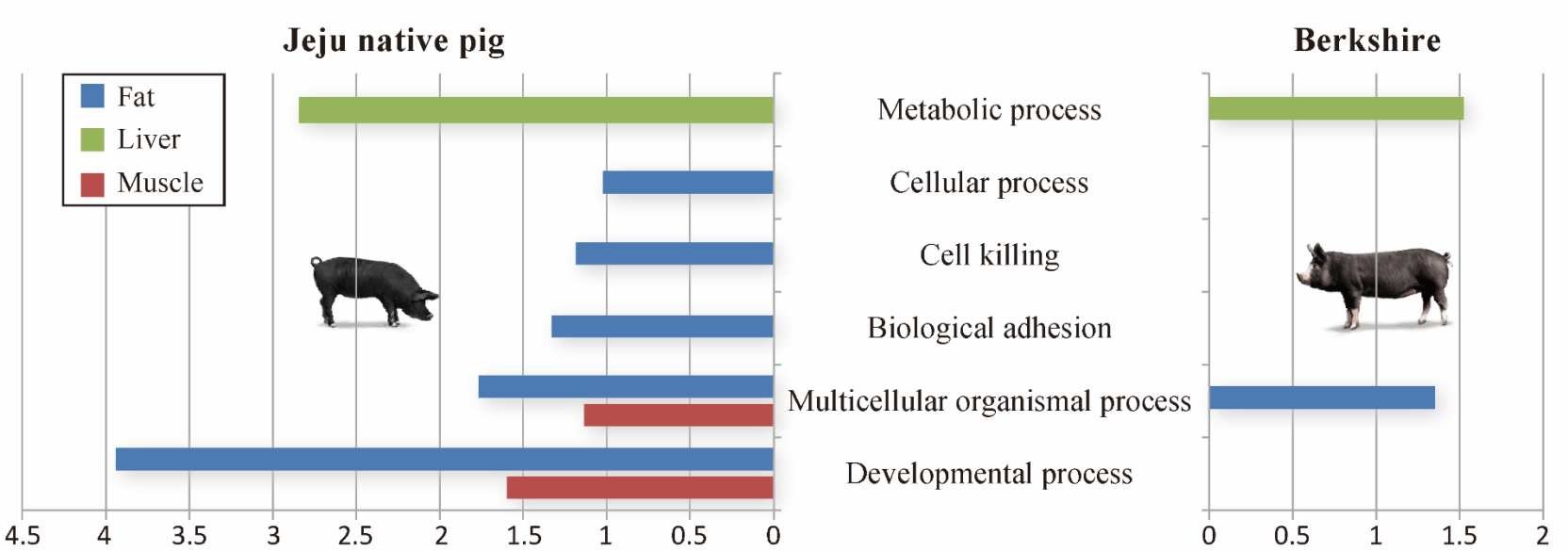


Figure 2.5. Up-regulation highest biological process GO terms of three tissues specific DEGs from JNP and Berks

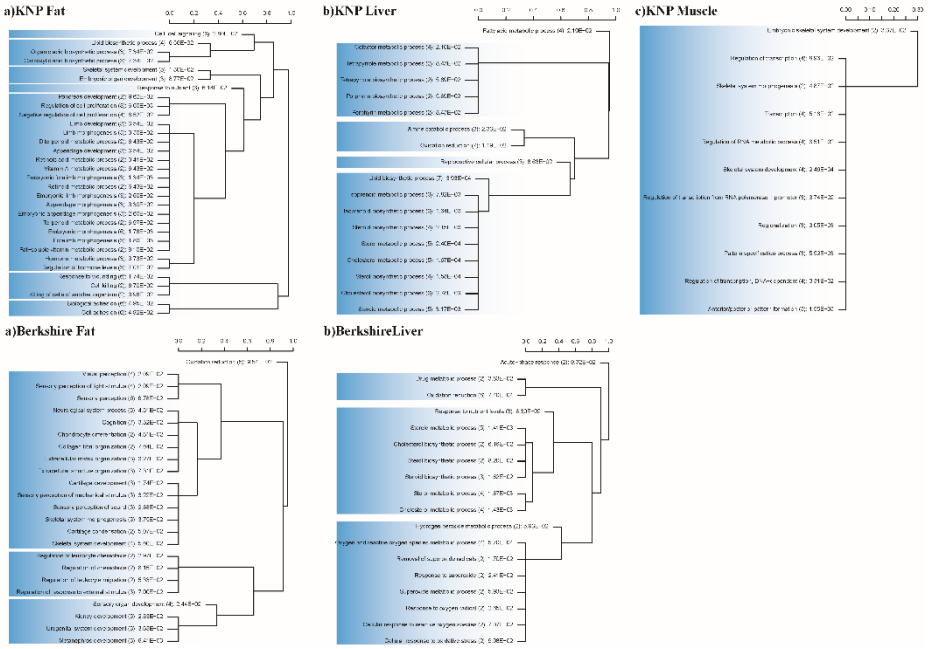


Figure 2.6. Up-regulation Biological process GO terms of three tissues specific DEGs from KNP and Berkshire

2.5 Discussion

I generated transcriptome data using Ilumina HiSeq2000 to investigate DEGs in three tissues of JNP and Berkshire pigs. These transcriptome data consist of highly reliable sequence reads, because the average mapping rate that passed the quality control (QC) filtering and the read mapping rate by the Sus scrofa (version 10.2) reference genome was >93% (Table 2.1). Based on this result, I identified DEGs in three tissues between JNP and Berkshire pigs, with a focus on genes related to meat quality and body growth.

Based on the functional annotation analysis that the gene ontology biological process terms, metabolic processes in liver and multicellular organismal processes in fat were simultaneously up-regulated in JNP and Berkshire pigs. In contrast, cellular processes, cell killing, biological adhesion and developmental processes in fat were up-regulated only in JNP. In addition, multicellular organismal and developmental processes in muscle were up-regulated only in JNP (Figure 2.4). Notably, the genes up-regulated in JNP (biological processes, cellular processes, biological adhesion and developmental processes in fat and developmental processes in muscle) were directly or indirectly related to meat quality and body growth. Moreover, the detailed gene ontology biological process terms (Figure 2.6) revealed significant up-regulation of cell-cell signaling lipid biosynthetic, limb development, hormone metabolic and fat-soluble vitamin genes in JNP fat, and skeletal system in JNP muscle; these have been associated with meat quality and body growth (Mc Meekan 1940, Corino et al. 1999, Kouba et al.

2003, Rehfeldt et al. 2006, Kim et al. 2009). In contrast, cartilage development and skeletal system development were up-regulated in the fat tissue of Berkshire pigs; these have been associated with body growth.

I found 26 genes related to meat quality and body growth in previous reports (Table 2.3). Most of these genes were up-regulated in Berkshire pigs. Among these genes, collagen genes (*COL11A1*, *COL11A2* (fat) and *COL2A1* (muscle)) are commonly related to cartilage development. As stated above, cartilage development constitutes an important mechanism for body growth through structural templates and the formation of most bones (Kim et al. 2007). This may provide evidence that Berkshire body growth is more effective than JNP. The average body weight of Berkshire pigs is 250–300 kg, which is greater than JNP. *MPZ* is known to function as a major structural protein of peripheral myelin, but is restricted to Schwann cells (Wagenknecht et al. 2005). MPZ is related to Dejerine–Sottas syndrome and congenital hypomyelination, and symptoms include slowly progressive distal muscle atrophy and weakness neuropathy (Warner et al. 1996). However, MPZ expression was up-regulated in JNP. As a result, MPZ may also be associated with the low body weight of JNP. *AADAT* expression was also up-regulated in JNP. *AADAT* in human is involved in lysine reduction, lysine biosynthesis and tryptophan metabolism, with the highest expression in the liver (Goh et al. 2002). It is important to mention that previous studies have demonstrated a strong relationship between lysine reduction and decreases in body weight and tissue protein; as a result, an increase in *AADAT* expression can interrupt growth mechanisms (Tesseraud et al. 1996). *LEPR* is highly expressed in liver, muscle, and intramuscular fat (IMF) (Muñoz et al. 2011). IMF is an

important factor for meat quality and is associated with marbling, juiciness, tenderness and flavor in pig (Cameron 1990). and there is a strong negative phenotypic correlation between IMF and moisture (Shi-Zheng et al. 2009, Gjerlaug-Enger et al. 2010). *LEPR* expression was down-regulated in JNP tissues. As a result, *LEPR* may have been associated with the high moisture content of JNP compared to Berkshire. *IGFN1* and *MYBPH* are related to both myosin in muscle, and to one another. *MYBPH* is located at crossbridge-containing C zones of striated muscle sarcomere. This is composed of globular *IGFN1*. In addition, myofibrils contain actin and myosin as the predominant protein in thin and thick filaments (Klont et al. 1998, Zhang 2009). The important term here is myosin, which is the major protein that comprises the thick filament and affects the muscle fiber type (MHC: myosin heavy chain). The muscle fiber type is reported to be related to the color, stability and tenderness of beef, and water content, color and eating quality of pork. I found that *IGFN1* and *MYBPH* were up-regulated in JNP. As a result, *IGFN1* and *MYBPH* may have affected the meat quality of JNP.

In addition, other DEGs are related to meat quality and body growth, but I cannot define clear correlations between other DEGs and JNP from the existing data. Therefore, in future studies, I will try to verify the correlations between other DEGs and JNP

This chapter will be published in elsewhere
as a partial fulfillment of Woncheoul Park's Ph.D program.

Chapter 3. RNA-seq analysis in the kidney of broiler chickens fed with diets containing different concentrations of calcium.

3.1 Abstract

Calcium (Ca) is an essential mineral required for the normal growth of life organisms. Ca plays an important role in cellular physiology, signal transduction, and mineralization of bone. In human, intakes of Ca lower than adequate level causes hypocalcemia and excessive intake of Ca causes hypercalcemia. In chicken, apart from mineralization of bone, Ca is required for the body weight gain and the formation of eggshell. However, genes that altered by low/high Ca intake, and that affect the body weight is not clear. In this study, I performed RNA sequencing (RNA-seq) in the kidney of broiler chickens fed with diets containing 0.8, 1.0 and 1.2 percent Ca. Because the kidney is one of the important organs involved in Ca homeostasis. By annotation of the RNA-seq data, I identified a significant number of differentially expressed genes (DEGs) in the kidney by a pairwise comparison manner using two tools such as cufflinks and edgeR. Using cufflinks, I identified 128 DEGs between 0.8 and 1.0 percent Ca, 141 DEGs between 0.8 and 1.2 percent Ca, and 103 DEGs between 1.0 and 1.2 percent Ca. Using edgeR, I identified 12 DEGs, of these a more strict and reliable 7 DEGs were overlapped with cufflinks. About 7 DEGs were further validated by real-time qPCR in the Ca supplemented kidneys, and the results were highly correlated with RNA-seq data. Next, the cufflinks/edgeR detected DEGs were subjected for the pathway enrichment, protein/protein interaction, and co-occurrence analysis to trace their involvement in diseases. My findings collectively suggest that higher (1.2 percent) intake of Ca than required amount could

reduce the body weight gain in broilers, and that affected DEGs were related to stress-induced disease such as hypertension.

3.2 Introduction

Calcium (Ca) is an essential mineral for normal cellular physiology and signal transduction in the life organisms. It is chiefly used for the mineralization of endoskeleton in higher vertebrates and exoskeleton in invertebrates. In the human body, 98 percent of Ca is stored in the skeleton, and only 2 percent is released into the extracellular fluids as calcium ion (Ca2+). The calcium ion is transported through bloodstream either as free ion or bound with career proteins to the functional site. The Ca is required for the formation of bone and teeth in vertebrates, shell in invertebrates, and eggshell in hard-egg laying species. The Ca involves in a wide spectrum of functions such as, acting as a secondary messenger for neuromuscular signaling, contraction of heart and muscles, hormone secretion, and acts as a cofactor for blood coagulation. Studies reported that Ca has a direct affect to membrane-spanning Ca receptor that is coupled through G proteins to intracellular signaling, and this receptor has been detected in several tissues including the parathyroid gland, kidney, brain, bone marrow, and breast (Riccardi et al. 1995, Gogusev et al. 1997, House et al. 1997, Rogers et al. 1997, Cheng et al. 1998).

Hypocalcemia and hypercalcemia are the clinical terms linked to abnormal Ca concentration in the blood. Hypocalcemia appears when Ca loss exceeds the normal level and it has two stages such as mild hypocalcemia and severe hypocalcemia. Hypercalcemia appears when Ca gain exceeds the normal level and it has three stages such as mild hypercalcemia, moderate hypercalcemia, and severe hypercalcemia (Goff et al. 1994, Bushinsky et al.

1998). The symptoms of severe hypocalcemia are associated with the loss/weakness of muscle and nerve function, milk fever in cows, and tetany in lactating cows, pigs and dogs. In addition, lower intake of Ca is related to both hypertension and preeclampsia (Morris et al. 1995, Levine et al. 1997, Tesfaye et al. 2015), and hypercalciuria is related to hypertension (Quereda et al. 1996). The Ca homeostasis is tightly controlled by two hormones called parathormone and calcitonin. The parathormone secreted from parathyroid gland increases the blood Ca level through resorption of Ca from bone, absorption of Ca in the intestine, and reabsorption of Ca in the kidney. In contrary, calcitonin secreted from the thyroid gland reduces the blood Ca level by inhibiting the Ca resorption from bone, and inhibiting the absorption/reabsorption of Ca in the intestine/kidney.

Adequate intake of Ca is indispensable to keep the healthy life for all organisms. However, Ca intake varies in human and animals depending on the situation and food habits. For example, high intake of Ca and vitamin D as hormone replacement to prevent bone loss is recommended for postmenopausal women. The vertebrate, carnivore and herbivore animals take more Ca than that of another essential mineral, phosphorus (P). On the contrary, seed-eating animals such as parrots and herbivora mice take more P than that of Ca. National Research Council (NRC) recommended the mineral requirements for several domestic animals. For example, the Ca requirement is about 1.2 percent for pre-ruminant calves, 1.5 percent for ruminant animals, and 1.5 to 1.9 percent for adult cows. About 2 percent Ca is recommended for all hind gut fermenters such as horse and rabbit (National Research Council 2005).

In the case of chicken, the NRC recommended Ca requirement vary according to the age and breeds [31]. The Ca requirement for 0 to 6 weeks (0.9%), 6 to 12 weeks (0.8%), and 12 to 18 weeks (0.8%) is same for the Leghorn-type white-egg-laying and brown-egg-laying strains. However, the Ca requirement for 18 weeks to first egg is 2.0 percent and 1.8 percent for the white-egg-laying strains and brown-egg-laying strains, respectively. Increasing the Ca intake ranging from 3 to 4.5g per laying hen per daily would be more beneficial for them. The old layers are needed to provide with high Ca to keep their eggshell strength. The Ca requirement for 0 to 3, 3 to 6, and 6 to 8 weeks of broiler chickens is 1.0, 0.9, and 0.8 percent, respectively (NRC. 1994). Studies reported that the Ca requirement is 1.5 percent for grower chicks, but 0.9 percent causes that reduced phytate digestion (Applegate et al. 2003). In addition, more than 2 percent of Ca intake leads to decrease in feed intake and weight gain, and increase of mortality rate (Fangauf et al. 1961). The maximum tolerable Ca intake for high-producing laying hens is 5 percent. Ca in layers diet is essential to make eggshell and to increase eggshell strength. Ca in broiler diet is essential for strengthening the bone, and increasing the productivity. Therefore, the accurate estimation of its requirement is important to maximize broiler productivity. The current NRC recommendations of Ca for optimal growth and bone formation in broiler chickens during 21-d posthatch is 1.0 percent (NRC. 1994). However, a few reports suggested that decreasing Ca level in diets improve growth performance of broiler chickens (Sebastian et al. 1996, Rao et al. 2006). This beneficial effect is associated with increased P utilization through the reduction of calcium phosphate formation in the intestinal tract (Selle et al.

2009), and decreased pH of the intestinal tract that favors digestive enzyme activity through reduction of buffering capacity (Walk et al. 2012). It was also reported that the higher intake of Ca may induce kidney malfunctions because poultry has a limited ability to handle high Ca loads in the blood (Collett 2012). In addition, 2 percent Ca in the diet of 7 days-old broiler causes the hypophosphatemia and decreases the growth rate (Hurwitz et al. 1995).

The kidney is one of the important organs involved in Ca homeostasis, and there might be several genes expressed in the kidney to support this function. However, there is no information regarding the genome-wide analysis in the chicken kidney in response to Ca supplementations. In this study, I generated RNA sequencing (RNA-seq) in the kidney of broiler chickens fed with diets containing three different concentrations of Ca (0.8, 1.0, and 1.2 percent). By quality screening and annotation of the RNA-seq reads, I identified several differentially expressed genes (DEGs) in the kidney samples between 0.8 and 1.0 percent Ca intake, 0.8 and 1.2 percent Ca intake, and 1.0 and 1.2 percent Ca intake, and the expression of seven candidate genes were analyzed by quantitative real-time PCR (qRT-PCR). Furthermore, by pathway analysis, interaction analysis and co-occurrence analysis, I identified DEGs that related to reduced weight gain, and identified that oxidative stress such as hypertension was associated with the reduced weight gain.

\

3.3 Materials and methods

3.3.1 Ethics statement and experimental design in body weight gain

The protocol for this experiment was reviewed and approved by the Institutional Animal Care and Use Committee at Chung-Ang University (IACUC No.: 14-0005). A total of 1,280 1-d-old Ross 308 broiler chicks (initial body weight (BW) = 39.4 ± 0.17 g) were used and were allotted to 1 of 3 dietary treatments with 6 replicates, each replicate consisting of 70 birds, in a completely randomized design. Chicks were obtained from a local hatchery (Yangji hatchery, Pyeongtaek, Republic of Korea) and were housed in conventional floor pens (200 cm × 230 cm × 100 cm = width × length × height for each pen) for 21 d. Three commercial-type experimental diets were formulated and the concentrations of Ca in 3 diets were 0.8, 1.0, and 1.2 percent each. The concentrations of non-phytate phosphorus (NPP) in all diets were maintained at 0.35 percent, and commercial phytase (Phyzyme XP, Danisco Animal Nutrition, Marlborough, UK) was supplemented to all diets at the level of 1,000 FTU/kg. All diets were formulated to meet or exceed the NRC (1994) requirements for broiler chickens during 21-d posthatch, with the exception of Ca and NPP. The diets were fed in mash form. All birds were provided with diets and water ad libitum. The room temperature was maintained at 30℃ during the first wk and then gradually decreased to 24℃ at the end of the experiment. A 24-h photoperiod was used throughout the experiment. The BW gain (BWG) and feed intake (FI) were recorded at the end of the experiment. Feed efficiency (G:F, g/kg) was calculated by dividing

BWG with FI. At the end of the experiment (21-d posthatch), 4 birds per treatment with a BW close to the treatment mean BW were euthanized by CO2 asphyxiation and immediately dissected. The kidney samples were collected, frozen with liquid-N, and kept in a freezer at -50℃ before further analysis.

3.3.2 RNA-seq library preparation and sequencing

RNA was isolated from 50~100 mg of the fragmented frozen kidney tissue samples from chicken broilers. TRIzolTM (Invitrogen, USA) reagent was used for the isolation of RNA. Tissue samples were homogenized in 1 ml of TRIzol reagent and 0.2 ml chloroform, which were subsequently precipitated by using 0.5 ml of 100% isopropanol. Isolated RNA samples were stored at -70℃. Total RNAs from kidney tissues were isolated from 10 chicken broilers, flash frozen on dry ice, and RNA was isolated using TRIzol™ (Invitrogen) reagent. Total RNA has been taken from the each sample for the construction and sequencing of the total RNA-seq library. The TruSeq RNA Sample Pre Kit was used according to the manufacturer's guidelines. Agilent Technologies Human UHR total RNA has been used as a positive control sample. The library was constructed according to a standard protocol provided by Ilumina, Inc. Libraries with different indexes were pooled together and were sequenced in one lane using an Ilumina HiSeq2000 high-throughput sequencing instrument with 100 pair-end (PE) reads.

3.3.3 Aligning raw reads to the chicken transcriptome

I trimmed the adapt sequence, the specific sequence of the other ILLUMINA and below 80bp reads by Trimmomatic ver 0.32 tool (Bolger et al. 2014) before alignment. After then, I aligned the transcript reads to the chicken (Gallus gallus) reference genome in ENSEMBL website (ftp://ftp.ensembl.org/pub/release-85/fasta/gallus_gallus/dna/) using HISAT2 ver 2.0.4 tool (Kim et al. 2015) that is a fast and sensitive alignment tool for mapping next-generation sequencing reads (both DNA and RNA). When this tool is used, I used the default option and added the option: --dta-cufflinks that report alignments tailored specifically for Cufflinks. Next, I used Featurecount tool (Liao et al. 2014) for counting the read in gene

3.3.4 Differentially expressed genes analysis

DEGs were identified using both the Cufflinks (Trapnell et al. 2012) and the R package edgeR (Robinson et al. 2010). First, I used the Cufflinks ver 2.2.1 tool following the default option, and added the option: -g/--GTF-guide that finds novel genes and isoform by Reference Annotation Based Transcript (RABT) assembly. Next, cuffdiff within cufflinks uses the normalized RNA-seq fragment to estimate the abundances of transcripts. Fragments per kilobase of exon per Million fragments mapped (FPKM) of each sample were counted to estimate the expression levels of the transcripts. Cuffdiff to estimate the differential expression of transcripts across condition points and finds significant changes in gene expression. Since Cuffdiff is able to detect

DEGs between two samples, I compared all calcium intake in pairs such as between 0.8 and 1.0, 0.8 and 1.2, and 1.0 and 1.2 percent. Significant DEGs were identified by false discovery rate (FDR) < 0.05. Second, I used the R package edgeR that is based on a negative binomial model and count data. When a negative binomial model is used, the dispersion should be computed before DEGs analysis is implemented. Generalized linear models (GLM) state probability distributions according to the relationship between mean and variance. The GLM likelihood ratio test is based on the idea of fitting negative binomial GLMs with Cox-Reid dispersion estimates. This automatically takes all known sources of variation into account. Significant DEGs were detected with a cutoff value of FDR < 0.1.

3.3.5 Pathway enrichment analysis

The chicken Ensembl gene IDs were converted to official gene symbols by cross-matching to human Ensembl gene IDs and official gene symbols. The official gene symbols of human homologues of chicken genes were used for functional clustering and enrichment analyses using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al. 2003). The representation of functional groups in three pairwise comparison treatments such as between 0.8 and 1.0, 0.8 and 1.2, and 1.0 and 1.2 percent Ca intake relative to the whole genome was investigated using the Expression Analysis Systematic Explorer (EASE) tool (Hosack et al. 2003) within

DAVID. Pathway analysis for the DEGs was carried out by Kyoto Encyclopedia of Genes and Genomes (KEGG) tool. To identify enriched KEGG pathway functionally clustered genes were filtered by EASE value < 0.1 and were detected.

3.3.6 Protein/protein interaction network and Co-occurrence analysis

Text mining methods from the literature and disease levels were combined to screen between DEGs in calcium intake and keyword such as hypertension and blood pressure, which were carried out in the COREMINE (http://www.coremine.com/medical/) (de Leeuw et al. 2012) and IPAD databases (http://bioinfo.hsc.unt.edu/ipad/) (Zhang et al. 2012), respectively. The DEGs and the exact keywords ‘hypertension', ‘blood pressure’, ‘chicken’, ‘kidney’, ‘weight gain’. ‘oxidative stress’ and ‘calcium’ were input in COREMINE for co‑occurrence analysis, and the disease information among DEGs were mined in the IPAD database. In addition, I used other web-based interfaces that construct the PPI network publicly available STRING database (http://string-db.org) (Franceschini et al. 2013). When this tool is used, I included only proteins with at least one connection and only connections with medium confidence scores assigned by STRING (>=0.4)

3.3.7 Quantitative real-time PCR

qRT-PCR analyses of miRNAs were carried out with TaqMan microRNA assays (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's protocol. The reverse transcription of miRNAs from miRNA(4 ng) was performed with miRNA-specific stem-loop primers, 10 nmol dNTP mix, 2.6 U of RNase Inhibitor, 33.5 U of MultiScribe RT enzyme, and 1 × RT Buffer (Applied Biosystems). The reaction was performed using a PCR Thermal Cycler Veritia™ (Applied Biosystems, Foster City, Calif., U.S.A.) with samples incubated at 16℃ for 30 minutes, 42℃ for 30 minutes and 85℃ for 5 minutes. cDNA was amplified using TaqMan PreAmp Master Mix Kit(Applied Biosystems, Foster City, CA, USA) according to the manufacturer's protocol. (14 cycle). PCR was performed in a ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Foster City, Calif., U.S.A.) in 384-well microtiter plates using a final volume of 20ul. Reaction mix consisted of 10 μL 2 × TaqMan Fast Universal PCR Master Mix, No AmpErase UNG, 1 μL 0.2 μM TaqMan probe, 3 μL 1.5 μM of forward primer, 1.4 μL 0.7 μM reverse primer, and 1.33 μL of cDNA. The PCR reactions were initiated with 10 minutes incubation at 95℃, followed by 40 cycles of 95℃ for 15 seconds and 60℃ for 60 seconds. All samples were amplified on triplicate and data were analyzed with Sequence Detector software (Applied Biosystems). cDNA was analyzed by BioRad CFX-96. All samples were measured in triplicate to ensure reproducibility, and Ct value was calculated using $2^{-2\Delta\Delta Ct}$ method (Livak et al. 2001).

Table 3.1. The primer sequence of DEGs used for qRT-PCR analysis

Gene	Accession No.	Primer sequence (5` to 3`)	
KLF2	ENSGALG00000003939	F	CCCACCTGCGGACACA
		R	CAGCCCTCCCAGTTGCA
HPX	ENSGALG00000022586	F	GCCGAGGGCACAGACAT
		R	ACTGCAGCGGTCACCAG
LECT2	ENSGALG00000006323	F	GATACGGCTGCGGCAATTAC
		R	GCCCTTGTGCTTTTCTCCTTT
NUP210	ENSGALG00000005078	F	GTCTCATCTCAAGGCAGCTAAAGTA
		R	GGTGGCAGACACTGGTAGAA
AP3S2	ENSGALG00000008291	F	CGGCTCGTCCGCTTCTAC
		R	CCGCAGCACCATAGAAGGT
ADAMTS8	ENSGALG00000001370	F	GCACTATGACACTGCCATCCT
		R	CGTGTCGCAGCCTTGATG
FABP4	ENSGALG00000015767	F	CCTGGAAGCTCCTTTCTAGTGAAAA
		R	CAGCCATCTTCCTGGTAGCAAA
GAPDH	ENSGALG00000014442	C	TCGTCAAGCTTGTTTCCTGGTATGA

3.4 Results

3.4.1 Growth performance as affected by three different Ca intakes

Chickens broiler fed diets containing 0.8 percent Ca had the greatest ($P<0.01$) body weight (BW), body weight gain (BWG), and feed intake (FI), whereas those fed diets containing 1.2 percent Ca had the least ($P<0.01$) BW, BWG, and FI among 3 different Ca intakes (Table 3.2). Broiler fed diets containing 0.8 or 1.0 percent Ca had greater ($P<0.01$) feed efficiency than those fed diets containing 1.2 percent Ca.

Table 3.2. Effect of dietary Ca concentrations on growth performance of broiler chickens during 21-d posthatch

	Dietary Ca concentrations			SEM	P-value
Items	1.2%	1.0%	0.8%		
Initial body weight, g	39	39	39		
21-d body weight, g	737[a]	855[b]	923[c]	18.4	<0.01
Body weight gain, g/21 d	697[a]	815[b]	884[c]	18.4	<0.01
Feed intake, g/ 21 d	1,036[a]	1,125[b]	1,201[c]	25.1	<0.01
Feed efficiency,[1] g/kg	673[a]	725[b]	736[b]	7.2	<0.01

[a,b,c] Means with different superscript differ at $P < 0.05$.

[1]Feed efficiency was calculated by dividing body weight gain (g) with feed intake (kg)

3.4.2 Quality of RNA-sequence reads among 3 different Ca intake

I acquired RNA-seq reads in 3 different Ca intake (0.8, 1.0 and 1.2 percent) from broiler kidney (GSE89544). Quality report revealed more than 94 percent reads with the average sequencing quality score passing Q30. The average numbers of sequence reads were 10.6, 11.1 and 11.7M in 0.8, 1.0 and 1.2 percent, respectively. Among the sequence reads that passed the Trimmomatic, on average were above 97 percent. In addition, most of alignment rate in 3 different Ca intake was above 95 percent which were mapped successfully to the chicken reference genome (*Galgal4*) using HISAT2. The numbers of total sequence reads, read order, index, yield and mapping rates for each sample are shown in Table 3.3. I used the several plotting methods such as dispersion, fpkmSCV, pairwise scatter, multi-dimensional scaling (MDS) and principal component analysis (PCA) plot for evaluating, clustering and exploring the quality of and the relationships between my 3 different Ca intake RNA-seq data in broiler kidney (Figure 3.1).

Table 3.3. Summary of RNA-seq reads and mapping rate of different calcium intake from ten chicken broiler individuals.

Lane	Sample ID	Read Order	Index	Yield(Bases)	# Reads	% of >= Q30 Bases(PF)	Passed-Trimmomatic	Overall Alignment rate
1	H_1_Kidney	1	CGATGT	1,183,345,896	11,716,296	94.15	11386200 (97.18%)	95.62%
1	H_1_Kidney	2	CGATGT	1,183,345,896	11,716,296	94.29		
1	M_1_Kidney	1	CCGTCC	1,002,354,199	9,924,299	93.77	9646488 (97.20%)	95.38%
1	M_1_Kidney	2	CCGTCC	1,002,354,199	9,924,299	94.26		
1	H_3_Kidney	1	ATGTCA	1,252,089,526	12,396,926	94.04	12026744 (97.01%)	95.29%
1	H_3_Kidney	2	ATGTCA	1,252,089,526	12,396,926	93.54		
1	M_2_Kidney	1	GTGAAA	1,131,942,148	11,207,348	94.10	10897037 (97.23%)	95.38%
1	M_2_Kidney	2	GTGAAA	1,131,942,148	11,207,348	94.41		
1	H_2_Kidney	1	CTTGTA	1,122,854,471	11,117,371	94.12	10798873 (97.14%)	95.49%
1	H_2_Kidney	2	CTTGTA	1,122,854,471	11,117,371	94.71		
2	L_1_Kidney	1	TAGCTT	1,013,764,270	10,037,270	94.19	9745171 (97.09%)	94.77%
2	L_1_Kidney	2	TAGCTT	1,013,764,270	10,037,270	94.57		
2	M_3_Kidney	1	TTAGGC	1,134,667,835	11,234,335	94.16	10917675 (97.18%)	92.96%
2	M_3_Kidney	2	TTAGGC	1,134,667,835	11,234,335	94.48		
2	L_4_Kidney	1	ATTCCT	1,130,748,429	11,195,529	94.25	10873936 (97.13%)	95.16%
2	L_4_Kidney	2	ATTCCT	1,130,748,429	11,195,529	94.87		
2	M_4_Kidney	1	GATCAG	1,201,298,343	11,894,043	94.29	11578895 (97.35%)	95.45%
2	M_4_Kidney	2	GATCAG	1,201,298,343	11,894,043	94.59		
2	L_3_Kidney	1	CGTACG	1,057,737,650	10,472,650	94.21	10173845 (97.15%)	95.12%
2	L_3_Kidney	2	CGTACG	1,057,737,650	10,472,650	94.03		

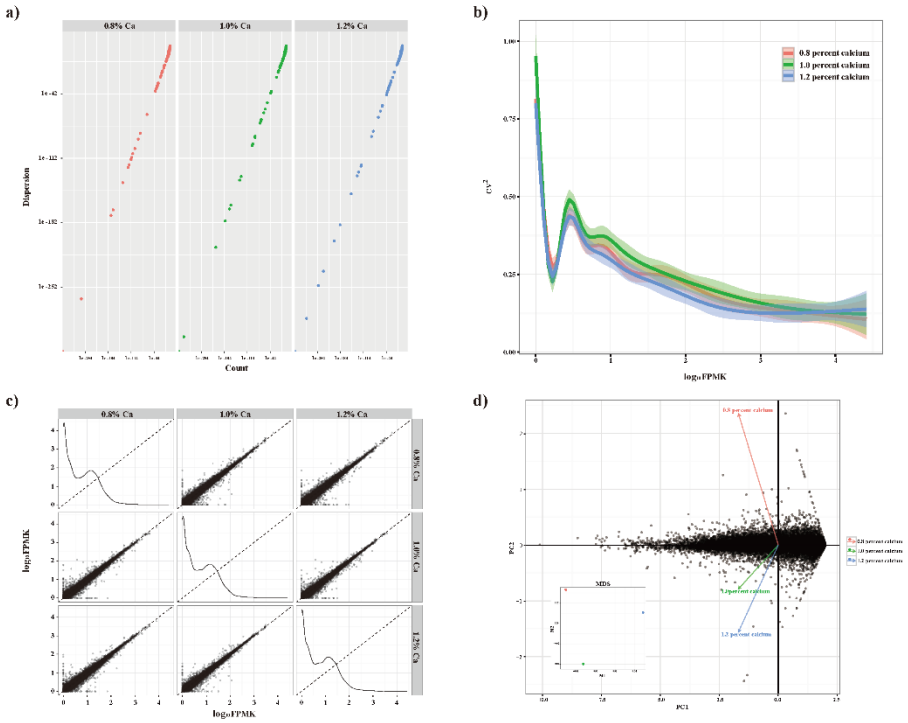


Figure 3.1. Summary of comparative analysis among three different calcium intake from RNA-seq data of kidney from ten chicken broiler (Total 9 samples: each of 0.8, 1.0 and 1.2 percent calcium intake is 3 samples).

a) Count vs dispersion plot by condition for all genes. b) The squared coefficient of variation allows visualization of cross-replicate variability between conditions and can be a useful metric in determining data quality at the gene level. c) Scatterplots for identify global changes and trends in gene expression between pairs of conditions. d) PCA plot and MDS plot for gene-level features.

3.4.3 Differentially expressed genes among 3 different Ca intake by both cuffdiff and edgeR

I identified DEGs using two programs such as cuffdiff and edgeR in 3 different Ca intakes such as 0.8, 1.0 and 1.2 percent from broiler kidney RNA-seq. First, I identified DEGs among 3 pairwise comparison treatments such as 0.8 and 1.0, 0.8 and 1.2, and 1.0 and 1.2 percent using cuffdiff within cufflinks (FDR < 0.05). As a result, the number of DEGs between 0.8 and 1.0 percent, 0.8 and 1.2 percent, and 1.0 and 1.2 percent were 128 (72 up-regulated, 47 down-regulated and 9 infinite value), 141(82 up-regulated, 45 down-regulated and 14 infinite value) and 103(58 up-regulated, 39 down-regulated and 6 infinite value), respectively. In addition, the number of common DEGs between 0.8 and 1.0 percent, 0.8 and 1.2 percent, and 1.0 and 1.2 percent was 25, 18 and 8, respectively. Moreover, the number of common DEGs in 3 pairwise comparison treatments was 1 such as Dipeptidyl peptidase Like 6 (*DPP6*) gene. Expression pattern of significant DEGs in 3 pairwise comparison treatments seems 4 patterns (Figure 3.2a, b and c). Second, I identified DEGs from test of the association as likelihood ratio test between 3 different Ca intakes and gene expression using GLM within edgeR (FDR < 0.1). As a result, total 12 DEGs were identified (5 up-regulated, 7 down-regulated), 7 genes such as Transmembrane Protein 8A (*TMEM8A*), Progastricsin (*PGC*), Hemopexin (HPX), Nucleoporin 210kDa (*NUP210*), Kruppel-Like Factor 2 (*KLF2*), Leukocyte Cell Derived Chemotaxin 2 (*LECT2*) and *GAL2* among 12 DEGs were overlap the DEGs that were identified using cuffdiff (Figure 3.2b and d, Figure 3.3 and Table 3.4). I performed qRT-PCR to validate the DEGs detected in broiler kidney. The 7

genes (*KLF2, HPX, LECT2, NUP210*: 4 common DEGs in both method such as cuffdiff and edgeR, and *AP3S2, ADAMTS8, FABP4*: 3 DEGs in edgeR) were randomly chosen (Table 3.1). The expression pattern of DEGs in RNA-seq were highly like qRT-PCR (Figure 3.4). Therefore, this result confirmed that DEGs in this study identified were reliable.

Table 3.4. Summary of DEG identified from the comparison among three different calcium intake using GLM within edgeR (FDR<0.1).

Ensembl ID	Human HGNC symbol	Chicken HGNC symbol	logFC	logCPM	LR	PValue	FDR	Association	Cufflinks overlap
ENSGALG00000003939	KLF2		-2.49671	8.031187	21.98473	2.75E-06	0.011116	Down regulated	8vs12, 10vs12
ENSGALG00000026188		TMEM8A	-4.43019	4.29225	20.94924	4.72E-06	0.014306	Down regulated	8vs10
ENSGALG00000001370	ADAMTS8	ADAMTS8	-2.1559	4.902834	19.12786	1.22E-05	0.029667	Down regulated	no
ENSGALG00000008291	AP3S2	AP3S2	-1.86177	5.509223	17.68768	2.6E-05	0.04551	Down regulated	no
ENSGALG00000028489	PGC		-3.33437	3.811141	17.6714	2.63E-05	0.04551	Down regulated	8vs10, 8vs12
ENSGALG00000022586	HPX	HPX	-2.94424	4.852415	16.20213	5.69E-05	0.069286	Down regulated	8vs10, 8vs12
ENSGALG00000028428			-5.30746	0.798146	15.37297	8.82E-05	0.089227	Down regulated	no
ENSGALG00000015767	FABP4	FABP4	9.578497	2.438523	23.73136	1.11E-06	0.00672	Up regulated	no
ENSGALG00000028627			9.545022	1.561843	24.04506	9.41E-07	0.00672	Up regulated	no
ENSGALG00000006323	LECT2	LECT2	3.637631	7.434093	16.70044	4.38E-05	0.06639	Up regulated	8vs12, 10vs12
ENSGALG00000005078	NUP210	NUP210	1.886229	6.587991	16.01596	6.28E-05	0.069286	Up regulated	8vs12
ENSGALG00000016669		GAL2	3.725195	5.388244	16.09272	6.03E-05	0.069286	Up regulated	8vs12, 10vs12

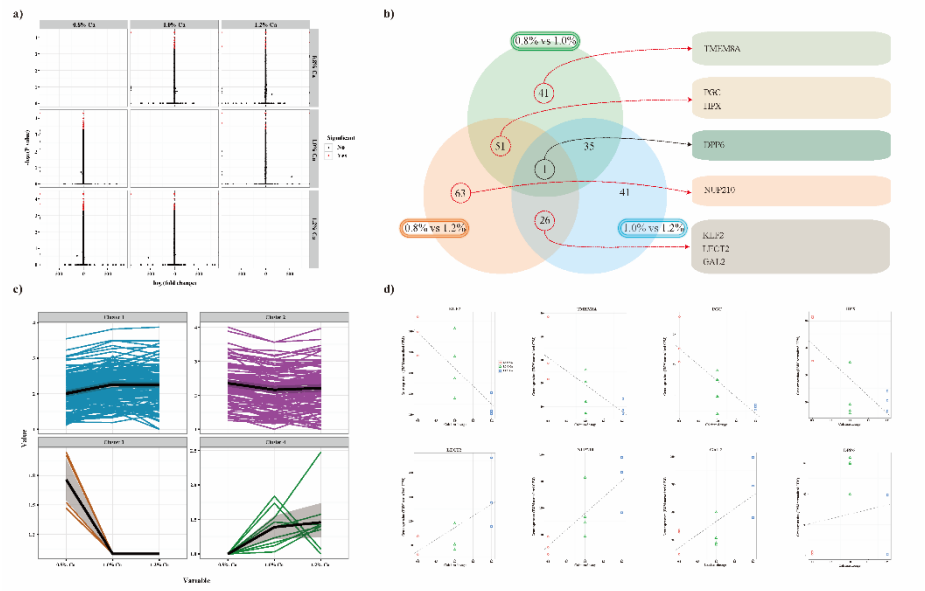


Figure 3.2. Identification of differentially expressed genes (DEGs) among three different calcium intake using both method such as cufflinks and edgeR.

a) Volcano plots explore the relationship between fold-change and significance. b) Venn diagram comparing DEGs between pairs of conditions (0.8 vs 1.0, 0.8 vs 1.2 and 1.0 vs 1.2 percent). Indicated in the diagram are the numbers of DEGs. c) Partitioning around medoids clustering with Jensen-Shannon distance for a CuffGeneSet. d) Scatterplots by 8 common DEGs that were identified by both cufflinks and GLM within edgeR

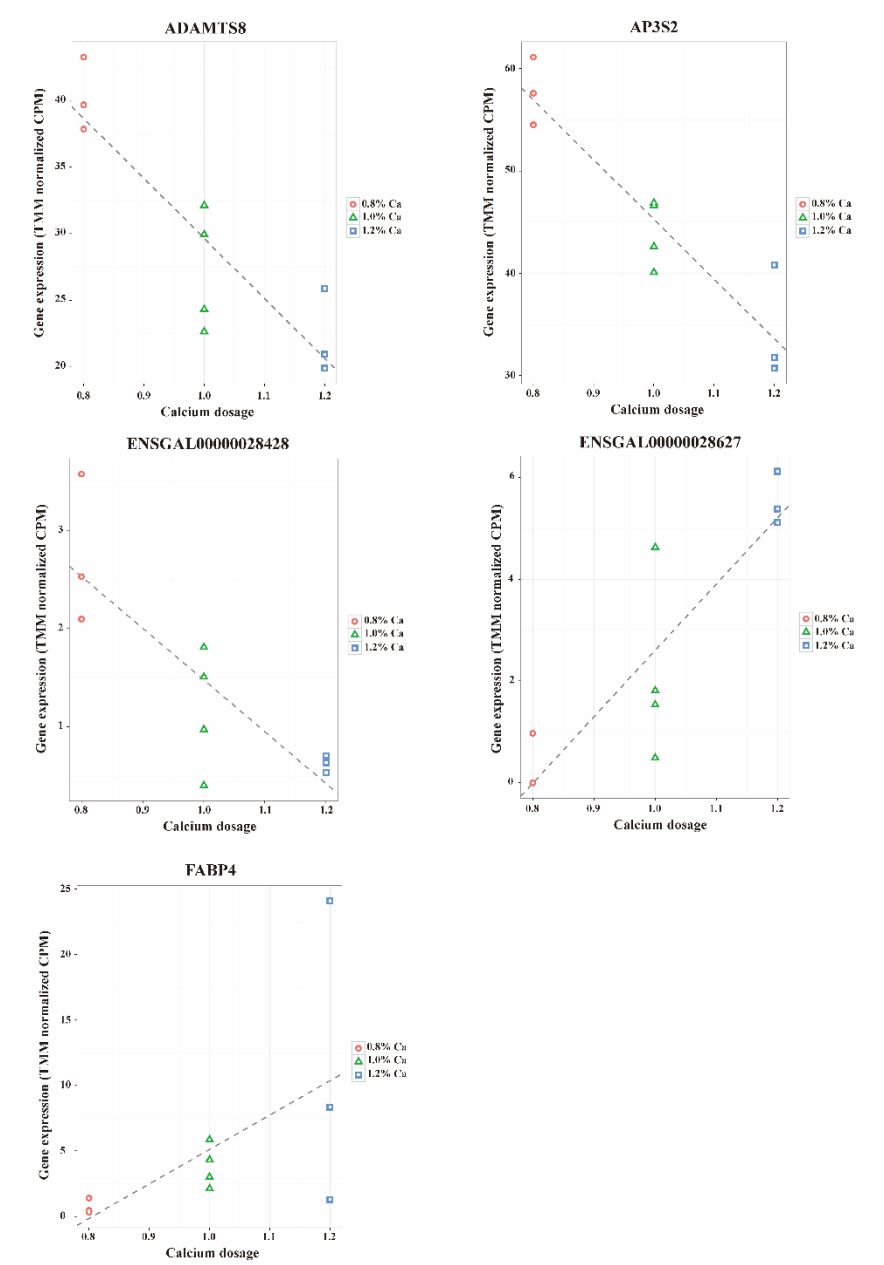


Figure 3.3. Scatterplots by 5 n DEGs that were identified by only GLM within edgeR.

2 up and 3 down regulated DEGs. (Up: ENSGAL00000028627, *FABP4* and Down: ENSGAL00000028428, *ADAMTS8*, *AP3S2*)

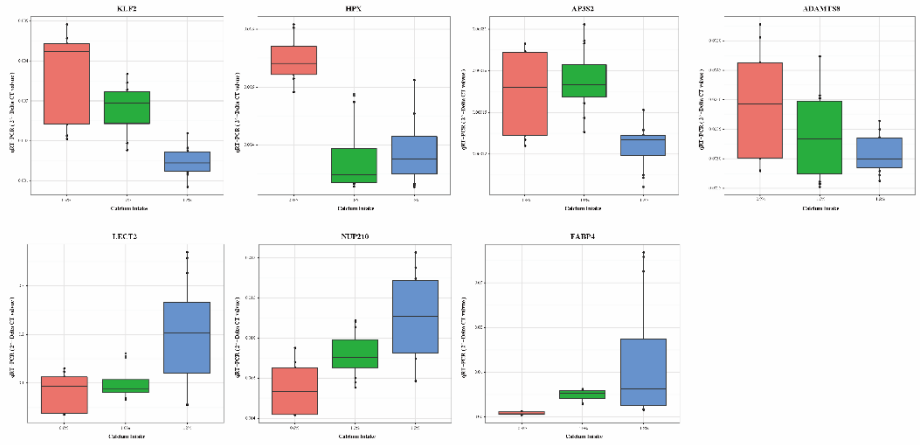


Figure 3.4. qRT-PCR validation of DEGs identified from the three different calcium intake RNA-seq data of chicken broiler kidney.

4 common DEGs in two method such as cuffdiff within cufflinks and GLM within edgeR, 3 DEGs in GLM within edgeR

3.4.4 Pathway enrichment, Protein/protein interaction network and Co-occurrence analysis

I used DAVID tool, STRING tool, COREMINE database and IPAD database for pathway enrichment, protein association networks and co-occurrence analysis. I used DEGs of 3 pairwise comparison treatments that were identified by cuffdiff, and then I analyzed KEGG pathway within DAVID tool. As a result, down regulated DEGs between 0.8 and 1.0 percent including the Metabolism of xenobiotics by cytochrome P450, Drug metabolism, Glutathione metabolism, Jak-STAT signaling pathway and Cytokine-cytokine receptor interaction, up regulated gene between 0.8 and 1.0 percent including the Toll-like receptor signaling pathway, Chemokine signaling pathway, Purine metabolism, Calcium signaling pathway, ECM-receptor interaction and Focal adhesion. down regulated DEGs between 0.8 and 1.2 percent including the Calcium signaling pathway, Neuroactive ligand-receptor interaction and Endocytosis, up regulated gene between 0.8 and 1.2 including the Progesterone-mediated oocyte maturation, Purine metabolism, Pyrimidine metabolism, Cell cycle, Oocyte meiosis and p53 signaling pathway (Figure 3.5a). But DEGs between 1.0 and 1.2 didn't exist the significant KEGG pathway, and DEGs from edgeR also didn't exist the significant KEGG pathway. Additionally, I check up on KEGG pathway of each DEGs from edgeR using KEGG pathway database, five genes such as *KLF2*, Adaptor Related Protein Complex 3 Sigma 2 Subunit (*AP3S2*), *NUP210*, *TMEM8A* and Fatty Acid Binding Protein 4 *(FABP4*) gene included FoxO signaling pathway, Lysosome, RNA transport, Neuroactive ligand-receptor interaction and PPAR signaling pathway (Regulation of lipolysis in adipocytes),

respectively (Table 3.5). I focused 12 DEGs from edgeR, then we perform STRING, COREMINE database and IPAD database analysis. To explore the interaction of the proteins translated by those 5 up and 7 down regulated DEGs in all DEGs from cuffdiff and edgeR, I used STRING that Protein/protein interaction network analysis. As a result, I found that predicted 4 up and 6 down regulated DEGs had a relationship between each other. In addition, predicted 4 up regulated DEGs were more correlated with each other than 6 down regulated DEGs (Figure 3.5b and Figure 3.6). The involvement of 8 DEGs (*HPX, PGC, ADMTS8, LECT2, NUP210, AP3S2, KLF2* and *FABP4*) in hypertension, 5 DEGs (*HPX, PGC, ADMTS8, KLF2* and *FABP4*) in blood pressure, 4 DEGs *(HPX, PGC, KLF2* and *FABP4*) in oxidative stress and 3 DEGs (*HPX, PGC* and *FABP4*) in weight gain determined in the COREMINE database, on the basis of literature co-occurrence, additionally *TMEM8A* gene only including both chicken and kidney. Moreover, chicken, kidney, calcium, hypertension, blood pressure, oxidative stress and weight gain were a lot of the basis of literature co-occurrence with each other (Figure 3.5c). The disease information in 12 DEGs were mined in the IPAD database. As a result, common disease in only 8 DEGs was a list of 54 diseases that included hypertension, drug toxicity and kidney diseases (Table 3.6).

Table 3.5. Enriched KEGG pathways from each of the DEGs that were identified by GLM within edgeR

Gene symbol	Gene full name	KEGG Pathway 1	KEGG Pathway 2
KLF2	Krueppel-like factor 2	FoxO signaling pathway	Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways
ADAMTS8	ADAM metallopeptidase with thrombospondin type 1 motif, 8	Degradation of the extracellular matrix	O-linked glycosylation
AP3S2	adaptor related protein complex 3 sigma 2 subunit	Lysosome	
PGC	progastricsin (pepsinogen C)		
HPX	hemopexin		
FABP4	fatty acid-binding protein 4	PPAR signaling pathway	Regulation of lipolysis in adipocytes
LECT2	leukocyte cell derived chemotaxin 2		
NUP210	nucleoporin 210kDa	RNA transport	
GAL2	galanin receptor 2	Neuroactive ligand-receptor interaction	
TMEM8A	transmembrane protein 8A		
ENSGALG00000028428	novel gene		
ENSGALG00000028627	novel gene		

Table 3.6. Enrichment analysis in the IPAD database from predicted 10 DEGs that were identified by GLM within edgeR

DiseaseID	DiseaseName	Molecule	AE	RE	N	MJI	Pvalue
MESH:D056486	Drug-Induced Liver Injury	PGC;LECT2;NUP210;HPX;FABP4;KLF2;ADAMTS8;AP3S2	8	1.17	15300	0.5003	0.691168
MESH:D009336	Necrosis	AP3S2;ADAMTS8;FABP4;KLF2;PGC;NUP210;LECT2;HPX	8	1.21	14853	0.5003	0.691168
MESH:D006973	Hypertension	NUP210;LECT2;PGC;FABP4;HPX;ADAMTS8;KLF2;AP3S2	8	1.26	14280	0.5003	0.691168
MESH:D004362	Drug Toxicity	AP3S2;ADAMTS8;KLF2;FABP4;HPX;NUP210;LECT2;PGC	8	1.27	14107	0.5003	0.691168
MESH:D005234	Fatty Liver	AP3S2;ADAMTS8;KLF2;HPX;FABP4;LECT2;PGC;NUP210	8	1.27	14162	0.5003	0.691168
MESH:D008107	Liver Diseases	FABP4;HPX;PGC;LECT2;NUP210;KLF2;ADAMTS8;AP3S2	8	1.27	14090	0.5003	0.691168
MESH:D007674	Kidney Diseases	FABP4;HPX;NUP210;PGC;LECT2;ADAMTS8;KLF2;AP3S2	8	1.28	14039	0.5003	0.691168
MESH:D006965	Hyperplasia	HPX;FABP4;LECT2;PGC;NUP210;KLF2;ADAMTS8;AP3S2	8	1.28	13955	0.5003	0.691168
MESH:D010146	Pain	KLF2;FABP4;ADAMTS8;AP3S2;LECT2;NUP210;PGC;HPX	8	1.28	13977	0.5003	0.691168
MESH:D002779	Cholestasis	ADAMTS8;AP3S2;KLF2;NUP210;PGC;LECT2;FABP4;HPX	8	1.3	13836	0.5003	0.691168
MESH:D006528	Carcinoma, Hepatocellular	ADAMTS8;AP3S2;KLF2;PGC;LECT2;NUP210;HPX;FABP4	8	1.3	13821	0.5003	0.691168
MESH:D009325	Nausea	HPX;LECT2;NUP210;PGC;KLF2;FABP4;ADAMTS8;AP3S2	8	1.3	13826	0.5003	0.691168
MESH:D011230	Precancerous Conditions	LECT2;PGC;HPX;NUP210;AP3S2;FABP4;KLF2;ADAMTS8	8	1.3	13753	0.5003	0.691168
MESH:D006470	Hemorrhage	KLF2;ADAMTS8;AP3S2;PGC;LECT2;NUP210;HPX;FABP4	8	1.32	13601	0.5003	0.691168
MESH:D006261	Headache	FABP4;HPX;PGC;LECT2;NUP210;KLF2;ADAMTS8;AP3S2	8	1.32	13561	0.5003	0.691168
MESH:D004487	Edema	FABP4;HPX;LECT2;PGC;NUP210;KLF2;AP3S2;ADAMTS8	8	1.33	13464	0.5003	0.691168

MESH:D009422	Nervous System Diseases	PGC;LECT2;NUP210;HPX;ADAMTS8;FABP4;KLF2;AP3S2	8	1.33	13526	0.5003	0.691168
MESH:D001943	Breast Neoplasms	AP3S2;ADAMTS8;KLF2;LECT2;PGC;NUP210;HPX;FABP4	8	1.34	13397	0.5003	0.691168
MESH:D051437	Renal Insufficiency	NUP210;HPX;PGC;LECT2;KLF2;FABP4;ADAMTS8;AP3S2	8	1.35	13241	0.5003	0.691168
MESH:D007249	Inflammation	KLF2;ADAMTS8;AP3S2;PGC;LECT2;NUP210;HPX;FABP4	8	1.35	13254	0.5003	0.691168
MESH:D001927	Brain Diseases	FABP4;HPX;LECT2;PGC;NUP210;ADAMTS8;AP3S2;KLF2	8	1.36	13134	0.5003	0.691168
MESH:D003072	Cognition Disorders	KLF2;ADAMTS8;AP3S2;NUP210;PGC;LECT2;FABP4;HPX	8	1.37	13131	0.5003	0.691168
MESH:D011507	Proteinuria	KLF2;FABP4;ADAMTS8;AP3S2;HPX;NUP210;PGC;LECT2	8	1.37	13045	0.5003	0.691168
MESH:D005334	Fever	NUP210;LECT2;PGC;FABP4;HPX;KLF2;ADAMTS8;AP3S2	8	1.39	12889	0.5003	0.691168
MESH:D009369	Neoplasms	FABP4;ADAMTS8;KLF2;AP3S2;HPX;PGC;NUP210;LECT2	8	1.39	12925	0.5003	0.691168
MESH:D000743	Anemia, Hemolytic	HPX;ADAMTS8;AP3S2;LECT2;PGC;NUP210;KLF2;FABP4	8	1.4	12764	0.5003	0.691168
MESH:D013375	Substance Withdrawal Syndrome	FABP4;ADAMTS8;KLF2;AP3S2;NUP210;HPX;LECT2;PGC	8	1.42	12665	0.5003	0.691168
MESH:D000230	Adenocarcinoma	NUP210;PGC;LECT2;FABP4;KLF2;HPX;AP3S2;ADAMTS8	8	1.43	12525	0.5003	0.691168
MESH:D058186	Acute Kidney Injury	NUP210;PGC;FABP4;LECT2;KLF2;HPX;ADAMTS8;AP3S2	8	1.43	12527	0.5003	0.691168
MESH:D003866	Depressive Disorder	FABP4;HPX;PGC;LECT2;NUP210;KLF2;ADAMTS8;AP3S2	8	1.43	12565	0.5003	0.691168
MESH:D007680	Kidney Neoplasms	ADAMTS8;KLF2;AP3S2;NUP210;LECT2;PGC;HPX;FABP4	8	1.47	12185	0.5003	0.691168
MESH:D009203	Myocardial Infarction	PGC;NUP210;LECT2;HPX;FABP4;ADAMTS8;KLF2;AP3S2	8	1.47	12232	0.5003	0.691168
MESH:D006330	Heart Defects, Congenital	KLF2;ADAMTS8;AP3S2;FABP4;HPX;NUP210;PGC;LECT2	8	1.48	12116	0.5003	0.691168
MESH:D009135	Muscular Diseases	FABP4;ADAMTS8;KLF2;AP3S2;NUP210;LECT2;PGC;HPX	8	1.48	12079	0.5003	0.691168
MESH:D002471	Cell Transformation, Neoplastic	PGC;LECT2;NUP210;HPX;FABP4;ADAMTS8;AP3S2;KLF2	8	1.49	12004	0.5003	0.691168

MESH:D014786	Vision Disorders	FABP4;AP3S2;ADAMTS8;KLF2;NUP210;HPX;LECT2;PGC	8	1.49	12066	0.5003	0.691168
MESH:D001523	Mental Disorders	FABP4;HPX;NUP210;PGC;LECT2;ADAMTS8;KLF2;AP3S2	8	1.5	11935	0.5003	0.691168
MESH:D006333	Heart Failure	KLF2;ADAMTS8;AP3S2;FABP4;HPX;NUP210;PGC;LECT2	8	1.51	11905	0.5003	0.691168
MESH:D013921	Thrombocytopenia	HPX;NUP210;LECT2;PGC;AP3S2;KLF2;FABP4;ADAMTS8	8	1.51	11870	0.5003	0.691168
MESH:D007565	Jaundice	AP3S2;KLF2;ADAMTS8;FABP4;HPX;PGC;LECT2;NUP210	8	1.51	11861	0.5003	0.691168
MESH:D011471	Prostatic Neoplasms	ADAMTS8;FABP4;KLF2;AP3S2;HPX;NUP210;LECT2;PGC	8	1.51	11873	0.5003	0.691168
MESH:D003875	Drug Eruptions	AP3S2;ADAMTS8;KLF2;FABP4;HPX;NUP210;LECT2;PGC	8	1.52	11793	0.5003	0.691168
MESH:D006331	Heart Diseases	LECT2;PGC;NUP210;HPX;FABP4;AP3S2;ADAMTS8;KLF2	8	1.52	11776	0.5003	0.691168
MESH:D008106	Liver Cirrhosis, Experimental	AP3S2;ADAMTS8;KLF2;NUP210;LECT2;PGC;HPX;FABP4	8	1.53	11682	0.5003	0.691168
MESH:D002318	Cardiovascular Diseases	AP3S2;ADAMTS8;KLF2;NUP210;PGC;LECT2;FABP4;HPX	8	1.56	11521	0.5003	0.691168
MESH:D004244	Dizziness	NUP210;PGC;LECT2;FABP4;HPX;KLF2;ADAMTS8;AP3S2	8	1.66	10813	0.5004	0.691168
MESH:D015430	Weight Gain	KLF2;ADAMTS8;FABP4;AP3S2;PGC;LECT2;HPX;NUP210	8	1.7	10561	0.5004	0.691168
MESH:D020246	Venous Thrombosis	HPX;NUP210;PGC;LECT2;AP3S2;FABP4;ADAMTS8;KLF2	8	1.71	10506	0.5004	0.691168
MESH:D050197	Atherosclerosis	AP3S2;ADAMTS8;KLF2;HPX;FABP4;PGC;LECT2;NUP210	8	1.77	10118	0.5004	0.691168
MESH:D009374	Neoplasms, Experimental	KLF2;ADAMTS8;FABP4;AP3S2;LECT2;NUP210;PGC;HPX	8	1.77	10153	0.5004	0.691168
MESH:D007319	Sleep Initiation and Maintenance Disorders	KLF2;FABP4;ADAMTS8;AP3S2;HPX;NUP210;PGC;LECT2	8	1.83	9784	0.5004	0.691168
MESH:D013610	Tachycardia	NUP210;HPX;PGC;LECT2;FABP4;ADAMTS8;KLF2;AP3S2	8	2.05	8753	0.5005	0.691168
MESH:D015428	Myocardial Reperfusion Injury	NUP210;LECT2;PGC;HPX;FABP4;ADAMTS8;KLF2;AP3S2	8	2.25	7963	0.5005	0.691168

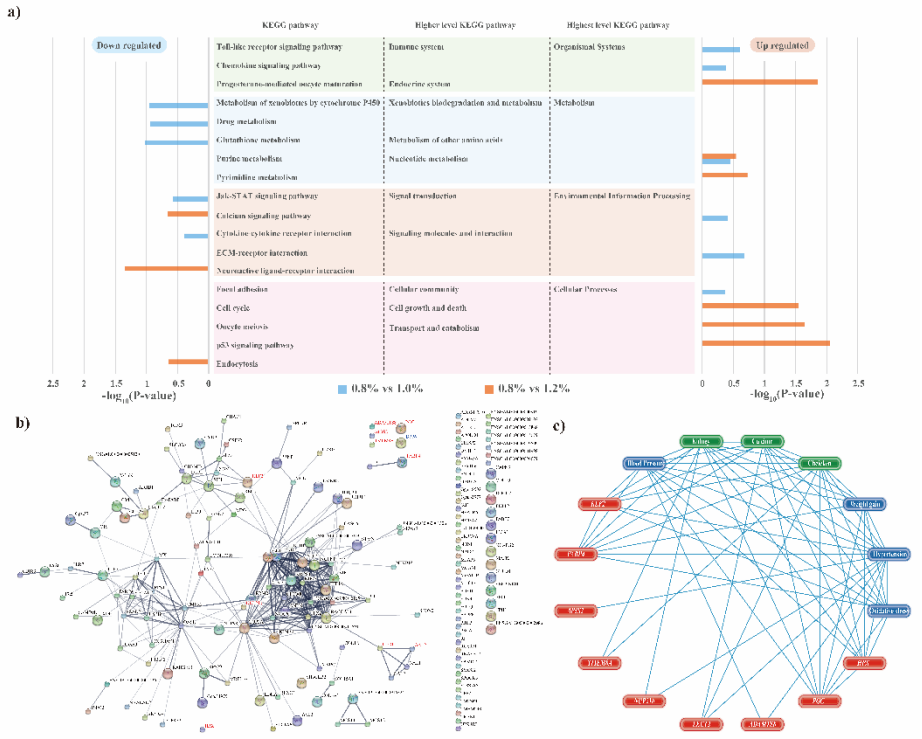


Figure 3.5. Enriched KEGG pathways, co-occurrence and Protein/protein interaction network analysis associated with DEGs.

a) For each set of up-regulated and down-regulated. DEGs between pairs of conditions (0.8 vs 1.0 and 0.8 vs 1.2), a KEGG pathway enrichment analysis was performed. b) Protein-protein interactions between total DEGs that were identified by cufflinks and GLM within edgeR as generated by the STRING database. c) Annotation of the co-occurrence of DEGs with blood pressure, hypertension, weight gain, oxidative stress, kidney, chicken and calcium using the Coremine Medical online database/tool

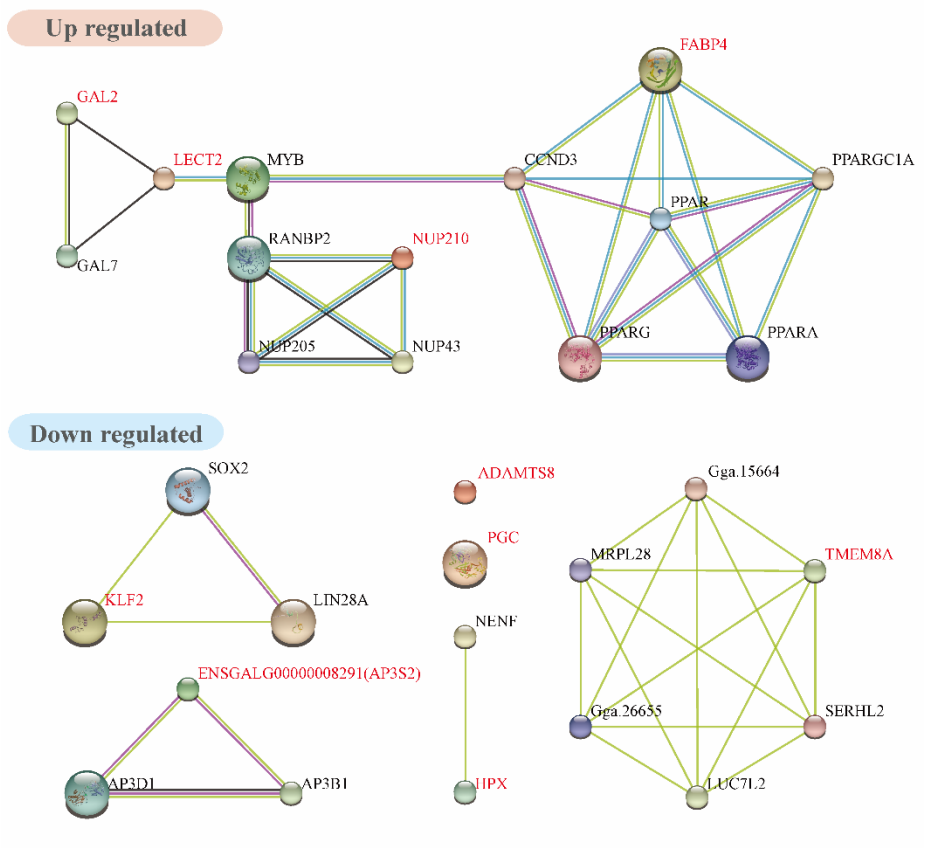


Figure 3.6. Protein/protein interaction network analysis associated with common DEGs between both tools such as cufflinks and GLM within edgeR.

4 up and 6 down regulated DEGs. Only proteins with at least 10 connection and only connections with medium confidence scores assigned by STRING (>=0.4)

3.5 Discussion

3.5.1 DEGs detection among Ca intakes from broiler kidney transcriptome using cuffdiff and edgeR

Ca is a chemical element and mineral in general feed additives of domestic animals. Effect and maximum tolerable Ca intakes in domestic animals were well known in previous studies and papers (Fangauf et al. 1961, Bushinsky et al. 1998, Power et al. 1999). Especially, these studies and paper were collected and decided the maximum tolerable levels of minerals such as Ca by NRC, and at that time NRC has been a benchmark publication with the maximum tolerable levels of minerals. However, poultry has been no more updated after 10 updated until 1994, since the first published in 1944 (Applegate et al. 2014). Therefore, I implemented the analysis of gene expression profile using RNA-seq through three different Ca intake in broiler kidney because I wants that the maximum tolerable Ca intakes should be reduced at the below level of current NRC recommendation. DEGs were identified by both cufflinks and the edgeR because I needed DEGs that were found not only in pairwise comparison and ordinal analysis in 3 Ca intakes, but also are more strict and reliable (Howard et al. 2013). As a result, in the present study, total 372 DEGs were identified between 0.8 and 1.0 percent, 0.8 and 1.2 percent, and 1.0 and 1.2 percent by using cuffdiff. In addition, I identified 12 DEGs using edgeR, then 7 DEGs (*TMEM8A, PGC, HPX, NUP210, KLF2, LECT2* and *GAL2*) were common gene by both methods, which can be a more strict and reliable DEGs. After then, I analyzed these

DEGs using pathway enrichment, protein association networks and co-occurrence analysis (Figure 3.5, Figure 3.6 and Table 3.4). Interpretation of the meaning from DEGs that were identified by cuffdiff was extremely difficult, but in KEGG pathway, I seem to have opposite tendency in Ca signaling between two treatments, for example, pathway 0.8 and 1.0 percent and 8 and 1.2 percent. This result suggests that 1.2 percent Ca intakes could have an adverse effect on broiler kidney (Figure 3.5a). By the result of all DEGs of STRING, I identified that linearly increased DEGs such as *FABP4, GAL2, LECT2* and *NUP210* genes were more hub and correlated genes than linearly decreased DEGs such as *KLF2, HPX, ADAMTS-8, TMEM8A, AP3S2* and *PGC* genes in 12 DEGs that were identified by edgeR (Figure 3.5b and Figure 3.3). I used IPAD database to confirm whether these 12 DEGs (10 predicted DEGs) are related to certain diseases and furthermore to check if diseases are associated with Ca. As a result, in total, 8 out of 10 DEGs were included diseases that related to brain, heart, kidney, liver and etc. Among these diseases, we have an interest in hypertension and weight gain because weight gain was an important factor in broiler industry and hypertension was related to stress-induced disease that has an impact on weight gain. However, P-value was no significant value (Table 3.6). The basis of literature co-occurrence between genes and keywords showed that most of the genes and keywords have related to papers but some genes and keywords have not (Figure 3.5c). Consequentially, I cannot give credence to COREMINE database, but I considered that COREMINE database has utilization in so many parts.

3.5.2 High Ca intakes lead to reduced weight gains and stress-induced disease such as hypertension

In my results, which seem to be decrease in the BW, BWG and FI by high Ca intakes. This observation agreed with Sebastian et al. (1996) who reported that high (1.25 percent) concentrations of Ca in diets had less BW and FI of broiler compared with the low (0.6 percent) or NRC recommended (1.0 percent) Ca intake. Rama Rao et al. (2006) also reported that decreasing Ca intake from 0.9 to 0.6 percent increased BW and FI of broiler. However, high Ca intakes adversely affects to the growth performance, and it can't be explained by experimental method alone because it is not possible to explain the molecular mechanisms. For this reason, I implemented the transcriptome analysis using broiler kidney RNA-seq. As a results, I found *FABP4*, *GAL2*, *LECT2*, *NUP210* and *DPP6* DEGs that expression level was linearly increased as the concentration of Ca increases (Figure 3.2d and Figure 3.3). *FABP4* gene was found in adipocytes and encodes the fatty acid binding protein. Fatty acid binding proteins are cytoplasmic proteins that unsaturated bind long-chain fatty acids and can reversibly bind hydrophobic ligands. In humans, increased plasma concentration of *FABP4* gene has been shown to be associated with atherosclerosis, cardiac diastolic dysfunction, hypertension, insulin resistance and obesity (Tso et al. 2007, Furuhashi et al. 2011, Ota et al. 2012, Fuseya et al. 2014). Enhanced expression of *FABP4* correlates with an increase in *CD36*, *CD68*, *CD52*, *CD163* and T-cell markers (Ayari 2015). This gene is actively released from human adipocytes in vitro via a non-classical, Ca-dependent mechanism as well as with coronary artery Ca (Bagheri et al. 2010, Schlottmann et al. 2014). This gene was related to hypertension and

expression level was higher in hypertensive patients than a normal person (Furuhashi et al. 2015). In addition, knockdown of this gene in deranged nutrient metabolism of diet-induced obese mice significantly increased BW and fat mass. In other words, this decreased gene expression level leads to increased BW and fat mass (Yang et al. 2011). And this increased gene expression level lead to decreased weight gain in epididymal adipose tissue of rats fed a fructose-rich diet such as an extract of chokeberry (Qin et al. 2012). *LECT2* gene was associated with adrenal amyloid and primary aldosteronism that have few or no symptoms such as muscular weakness, high blood pressure and headaches (Rauschecker et al. 2015, Wang et al. 2015). This gene was isolated like a neutrophil-chemotactic factor produce d by T cells (Yamagoe et al. 1996). In addition, this gene was related with β-catenin that plays an important role in tumorigenesis, and increased *LECT2* gene expression level was specific to tumors induced by β-catenin tumorigenesis. Moreover, activation of *LECT2* gene was reported to be leading to development of tumors with a better prognosis (Ovejero et al. 2004, Anson et al. 2012). In recent studies (2014 ~ 2016), this gene is regarded as recently discovered hepatokine that mediates obesity-related metabolic disturbances and insulin resistance (Lan et al. 2014, Hwang et al. 2015, Chikamoto et al. 2016). And hepatokine *LECT2* amyloidosis was related to portal hypertension (Damlaj et al. 2014). These studies make a general comment that increased *LECT2* expression level is commonly observed in insulin resistance and obesity in human and mice. However, the role of *LECT2* in the development of obesity and insulin resistance induced by over-nutrition has not yet been established. Nucleoporin 210kDa (*NUP210*), also known as Nuclear pore

glycoprotein-210 (*gp210*) was associated with diseases such as autoimmune disease of urogenital tract and primary biliary cirrhosis (PBC), and PBC was related to pulmonary hypertension (PH) and polymyositis (PM) (Nakamura et al. 2006, Honma et al. 2008). In addition, anti-gp210 antibodies have diagnosed PBC with jaundice and liver failure. (Ishibashi et al. 2011) *NUP210* gene was initially identified as an early response gene to cause of metanephric kidney development in Mouse (Olsson et al. 1999). *GAL2* gene was known that galactose permease and it's required for utilization of galactose, also able to transport glucose (Kasahara et al. 2000, Rodríguez et al. 2000, Maier et al. 2002). In addition, this *GAL2* gene was found to be the upregulated transcripts in whole blood cells of wild passerine, after immune with lipopolysaccharides (LPS) stimulation (Meitern et al. 2014). *DPP6* gene associated with ischemic heart disease (Makeeva et al. 2015). Therefore, I suggest that linearly increased *FABP4* and *LECT2* gene expression level as the concentration of Ca increases in broiler kidney could lead to more T-cells being activated, effected the calcium-dependent mechanism, play a protective role in tumorigenesis and may directly induced hypertension. Moreover, *FABP4* gene expression level negatively correlated with BW and fat mass, but *LECT2* gene expression level positively correlated with obesity and insulin resistance. Thus, I speculated that high Ca intakes in broiler were such that even if BWG and FI decreased, that broiler already has the obesity and high insulin resistance. However, I cannot corroborate others gene such as *NUP210*, *GAL2* and *DPP6* as for blood pressure and hypertension, but can suggest that these genes were indirectly related to blood pressure and hypertension.

I found that the expression level of *KLF2*, *HPX*, *ADAMTS-8*, *TMEM8A*, *AP3S2* and *PGC* genes was linearly decreased as the concentration of Ca increases (Figure 3.2d and Figure 3.3) *KLF2*, also known as lung Krüppel-like Factor (*LKLF*) gene is a member of the Krüppel-like factor family in broadly expressed zinc finger transcription factors, and it was associated with the lung development, embryonic erythropoiesis, epithelial integrity, T-cell viability, adipogenesis, B cell homeostasis, plasma cell homing and vascular growth/remodeling (Pearson et al. 2008, Winkelmann et al. 2011, Novodvorsky et al. 2014). The overexpression level of this gene in human and murine, number of proangiogenic cells (PACs) is increased by 60% in vitro, and neovascularization abilities of aged murine PACs is improved in an ischemic hind limb model in vivo, respectively (Boon et al. 2010). However, the number of PACs and neovascularization abilities was corrupted by risk factors for ischemic heart disease such as age, hypertension, or smoking (Vasa et al. 2001). This gene expression level in the developing chick heart was decreased by Trichloroethylene (TCE), which may function to alter endothelial development (Makwana et al. 2010). Moreover, this gene affected to wall shear stress that was related to blood flow in heart development, and this gene expression level was decreased in areas of low and disturbed wall shear stress (Groenendijk et al. 2007). Therefore, *KLF2* gene with increased expression level improves portal hypertension (Marrone et al. 2015). Moreover, this gene is an adipogenesis inhibitor that is related to the obesity, and increase this gene expression level by retinoic acid leads to prevent diet-induced weight gain (Berry et al. 2012). Hemopexin (*HPX*) gene is a haem binding protein and the structure of chicken HPX gene is more complex than

that of mammalian *HPX* gene. This gene is the acute phase response in chicken and expression level was increased in many infection (O'Reilly 2016). In addition, *HPX* gene expression level was decreased in idiopathic intracranial hypertension and preeclampsia that caused the symptoms such as hypertension, pitting edema, epigastric pain and swelling (Brettschneider et al. 2011). Moreover, this gene was not only associated with daily weight gain and backfat thickness in large white pig by protein phenotypes, but also associated with susceptibility or resistance to diet-induced obesity (Tagliaro et al. 1995, Choi et al. 2012). In addition, *TMEM8A* gene was also associated with preeclampsia (Enquobahrie et al. 2008). *ADAMTS-8* gene was a member of the ADAMTS protein family and aggrecanases. *ADAMTS-8* gene was identified as influencing pulse pressure and mean arterial pressure by genome-wide association study, and this gene expression level was decreased in brain tumors (Dunn et al. 2006, Wain et al. 2011, Wain 2014). Crucially, this gene related to hypertension has a patent application in 2007 (Salonen et al. 2007). *AP3S2* gene was associated with carotid plaque and obesity with type 2 diabetes mellitus (Dong et al. 2010, Dong et al. 2012, Sanghera et al. 2012, Shahid et al. 2016). *PGC*, also as known as Pepsinogen C gene is one of the two major groups of pepsinogen. This gene was associated with cancerous of the pancreas, atrophic gastritis and lung injury/disease (Borch et al. 1989, Grützmann et al. 2003, Sato et al. 2004, Fukushima et al. 2007, Ballard et al. 2010). Therefore, I suggest that linearly decreased expression level of *KLF2*, *HPX*, *TMEM8A*, *ADAMTS-8* and *AP3S2* genes in broiler kidney may directly or indirectly related to blood pressure with hypertension and weight gain.

However, I don't have a leg to stand on *PGC* gene, because I could not find clear evidence for the blood pressure with hypertension and weight gain.

Conclusions, I demonstrate the empirical result that concentration of Ca increase leads to reduced BWG and FI by using transcriptome analysis such as the pathway enrichment, protein association networks and co-occurrence analysis from DEGs that are found by using methods such as cuffdiff and edgeR. First, I identified DEGs that directly are related to weight gain. Second, I also identified DEGs that are related to stress-induced disease such as hypertension that affects weight gain. Although a few of DEGs have not been previously associated with blood pressure, hypertension and weight gain, I suggest that these DEGs may play a role in blood pressure with hypertension and weight gain, and further studies should carry out additional investigation of their roles and functions. These findings contribute to a better understanding of the molecular mechanisms potentially underlying correlation among Ca intakes, BWG, FI and stress-induced such as hypertension, and may provide important information relevant to other species, especially humans. Therefore, I do not support the fact that the existing maximum tolerable Ca intake of NRC in 2005 is 1.0 percent, but I suggest that the maximum tolerable Ca intake should be reduced at the below level of current NRC recommendation.

This chapter comprises a part of paper published in *PLOS ONE* as a partial fulfillment of Woncheoul Park's Ph.D program.

Chapter 4. Investigation of de novo unique differentially expressed genes related to evolution in exercise response during domestication in Thoroughbred race horses

4.1 Abstract

Previous studies of horse RNA-seq were performed by mapping sequence reads to the reference genome during transcriptome analysis. However in this study, I focused on two main ideas. First, differentially expressed genes (DEGs) were identified by de novo–based analysis (DBA) in RNA-seq data from six Thoroughbreds before and after exercise, here-after referred to as "de novo unique differentially expressed genes" (DUDEG). Second, by integrating both conventional DEGs and genes identified as being selected for during domestication of Thoroughbred and Jeju pony from whole genome re-sequencing (WGS) data, I gives a new concept to the definition of DEG. I identified 1,034 and 567 DUDEGs in skeletal muscle and blood, respectively. DUDEGs in skeletal muscle were significantly related to exercise-induced stress biological process gene ontology (BP-GO) terms: 'immune system process'; 'response to stimulus'; and, 'death' and a KEGG pathways: 'JAK-STAT signaling pathway'; 'MAPK signaling pathway'; 'regulation of actin cytoskeleton'; and, 'p53 signaling pathway'. In addition, I found *TIMELESS*, *EIF4A3* and *ZNF592* in blood and *CHMP4C* and *FOXO3* in skeletal muscle, to be in common between DUDEGs and selected genes identified by evolutionary statistics such as F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH). Moreover, in Thoroughbreds, three out of five genes (*CHMP4C*, *EIF4A3* and *FOXO3*) related to exercise response showed relatively low nucleotide diversity compared to the Jeju pony. DUDEGs are not only conceptually new DEGs that cannot be attained from reference-based analysis (RBA) but also supports previous RBA results related to exercise in

Thoroughbred. In summary, three exercise related genes which were selected for during domestication in the evolutionary history of Thoroughbred were identified as conceptually new DEGs in this study.

4.2 Introduction

Since domestication, at around 3500 B.C.E, horses have mainly been used for riding and racing(Weatherby 1791). One domesticated breed of horses, the Thoroughbred, has been specifically bred for speed, endurance, and strength since the 18th century. The extreme selection for these traits has resulted in a highly adapted athlete (Poole 2004) with very high aerobic capacity (Young et al. 2002), and high skeletal muscle mass (Kayar et al. 1989), which comprises over 55% of total body mass (Gunn 1987). The Thoroughbred is an excellent breed for competitive horse racing and by extension a valuable model for studying exercise response. A previous study has shown that exercise training in Thoroughbreds resulted in coordinated changes in the expression of genes related to metabolism, oxidative phosphorylation and muscle structure (McGivney et al. 2010).

Domestication leads to gradual changes at the genetic level by a process of selection in a population of animals or plants. Most domestic animals were selectively bred for the goal of benefitting human beings. Due to the combined effect of natural selection and human-controlled selective breeding, phenotypic changes, which are related to genetic mutation, accompany the domestication process. Some genetic mutations with beneficial phenotypic effects have been either highly enriched or vanished by selective sweeps (Andersson 2012). A selective sweep is the reduction of genetic diversity in the neighboring DNA of a fixed mutation. Selective sweep regions in the genome can potentially be identified by a genome scan, and the low variation

interval surrounding the selected gene can be found by fine-scale mapping. Using such genome scans, selective sweeps have been identified in domestic and natural (wild progenitor) populations (Storz 2005, Wright et al. 2005).

Previous horse transcriptome studies using RNA-seq were carried out by mapping sequence reads to a reference genome. However, reference genome assembly has been known to have flaws including missing expressed genes (Chen et al. 2011), hundreds to thousands of miss-assemblies and large genomic deletions (STEVEN et al. 2005), and problems in trans-spliced genes (Kinsella et al. 2011). Therefore, the results and success of reference transcriptome assembly depends on both the availability and quality of the reference genome. On the contrary, de novo transcript assembly has several advantages. First, it does not depend on a reference genome (Birzele et al. 2010). This is a key advantage as many organisms do not have a high-quality finished reference genome. For these organisms, de novo assembly becomes the first analysis step. Also, it does not depend on the correct alignment of reads to known splice sites (Burset et al. 2000) or the prediction of novel splicing sites, both of which are required by reference-based assemblers. Trans-spliced transcripts and similar transcripts originating from chromosomal rearrangements can be assembled using the de novo approach. In addition, de novo transcriptome assembly can help researchers investigate genes that are absent in the reference genome due to the incompleteness of reference sequences (Chen et al. 2011). Lastly, it can identify new transcripts and new transcript structures (Robertson et al. 2010, Chen et al. 2011). However, reconstruction of full-length transcripts from short reads with considerable sequencing error rates poses substantial computational

challenges (Grabherr et al. 2011). Still, de novo assembly in RNA-seq is an important approach for carrying out transcriptomic studies.

Recently, many de novo assembly software tools have been developed, most of which take the de Bruijn graph approach. This approach usually has two important parameters: k-mer length and coverage cutoff value (Chen et al. 2011). Tools such as Trans-ABySS (Robertson et al. 2010), Trinity (Grabherr et al. 2011), ABySS (Birol et al. 2009), Oases (Schulz et al. 2012), Rnnotator (Martin et al. 2010), Multiple‑k (Surget-Groba et al. 2010), SOAPdenovo (Li et al. 2009) and Velvet (Zerbino et al. 2008) follow this approach. Considering these de novo assembly software tools, Manfred G Grabherr, et.al concluded that Trinity de novo assembly software tool is superior to others for a number of reasons. Specifically, Trinity fully reconstructs a large fraction of transcripts, including alternatively spliced isoforms and transcripts from recently duplicated genes. In addition, Trinity resolves ~99% of the initial sequencing errors, determines splice isoforms, and distinguishes transcripts from recently duplicated and identified allelic variants(Grabherr et al. 2011).

4.3 Materials and methods

4.3.1 Ethics statement

This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of Pusan National University. All experimental procedures used in this study were approved by the Institutional Animal Care and Use Committee of the Pusan National University (PNU-2013-0417). The owners of the Thoroughbred horses gave permission for their animals to be used in this study.

4.3.2 Analysis of horses RNA-seq data

RNA-seq data between before and after exercise

I generated RNA-seq data from six Thoroughbred horses before and after exercise as described in a previous study (Park et al. 2012). Samples of skeletal muscle and blood were taken from six Thoroughbreds before and after exercise. ‘Before exercise’ samples were collected from the triceps brachii of the right leg and from the jugular vein and carotid artery of each horse. After an adequate resting period of several hours, the horses were subjected to a 30-min trot. Then, immediately after this trot, the ‘after exercise’ samples were collected from the same tissues of each individual. Thoroughbreds usually canter for 17-18 min per day. For the purposes of this study, a 30-min trot was taken to be the equivalent to 17-18 min of cantering.

Total RNA from the skeletal muscle and blood samples were isolated using TRIzol (Invitrogen) and the RNeasy RNA purification kit with DNase treatment (Qiagen). mRNA was isolated from the total RNA using oligo-dT beads, then reverse transcribed into double-stranded cDNA fragments. Construction and sequencing of an RNA sequence library for each sample was carried out based on Illumina HiSeq2000 protocols in order to generate 90bp paired-end reads. Twenty-four sets of transcriptome data were generated from muscle and blood samples of six horses obtained before and after exercise. The RNA-sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE37870.

De novo-based analysis (DBA)

I used the Trinity *de novo* assembly software tool (Grabherr et al. 2011, Henschel et al. 2012) following the default settings, except for the following options: number of CPU and alignment method: bowtie. First, Trinity tools generated a reference for each obtained sample (a total of 24 samples) to detect DEGs. Second, Trinity tools generated a reference for each individual (a total of 4 samples) to compare with the SNPs from the whole genome sequence, RNA-seq (using reference transcriptome assembly) and RNA-seq (using *de novo* assembly). The component ID was converted to known transcript ID (ftp://ftp.ensembl.org/pub/release-73/fasta/equus_caballus/cdna/) using Blastall (Tao 2006), a user-friendly, free open source tool, which is suited for short read alignment. After conversion, I filtered the transcript ID by alignment length of higher than 80 percent. Reference-based analysis

(RBA): Most of the RBA used in this study is described in Kim et al. (2013)' works. TopHat (Trapnell et al. 2009) (ver.1.4.1) was used to map the sequences to a horse reference genome and annotated using the EquCab2 database (http://hgdownload.cse.ucsc.edu/downloads.html#horse).

DEG selection (*de novo* vs reference)

I examined the differential expression of replicated count data by applying a method based on negative binomial model as implemented in the R package edgeR (Robinson et al. 2010). This package was used because RNA sequence data may exhibit more variability than expected in a Poisson distribution due to wide dispersal in the genome. The method implemented in the edgeR package automatically takes all known sources of variation into account. Significant DEGs were detected with a cut-off value of FDR <0.01, based on a paired design between 'before exercise' and 'after exercise'

Genotype calling and SNP calling (*de novo* vs reference)

Three open source packages were used for downstream processing and variant calling: Picard Tools (http://picard.sourceforge.net), SAMtools (Li et al. 2009) and the genome Analysis Toolkit (GATK (McKenna et al. 2010)). Substitution calls were made with the GATK Unified Genotyper (DePristo et al. 2011). All calls with a Phred-scaled quality of less than 30 were filtered out and VCFtools (Danecek et al. 2011) was used for handling the vcf file format.

DAVID analysis (*de novo* vs reference)

One simple but extremely widely used systems biology technique for highlighting biological processes is gene category over-representation analysis. In order to perform this analysis, genes are grouped into categories by a common biological property and then tested to find categories that are over represented among the differentially expressed genes. Gene Ontology (GO) categories are commonly used in this technique and there are many tools available for performing GO and KEGG pathway analysis. I used DAVID (Da Wei Huang et al. 2008) web tool to convert the equine Ensembl gene IDs to official gene symbols. This was carried out by cross-matching the equine Ensembl gene IDs to the human Ensembl gene IDs and the official gene symbols. The representation of functional groups in blood and skeletal muscle relative to the whole genome was investigated using the Expression Analysis Systematic Explorer (EASE) tool (Hosack et al. 2003) within DAVID. The EASE tool is a modified Fisher's exact test used to measure enrichment of gene ontology (GO) terms (Alterovitz et al. 2010). To identify enriched GO terms, functionally clustered genes were filtered by an EASE value of <0.01. In addition, A KEGG pathway enrichment test was performed using EASE, with a cut-off value <0.1

Quantitative real-time reverse transcript-PCR (qRT-PCR) validation

A blood sample was obtained from a Thoroughbred horse maintained at Ham-a Racing Horse Resort & Training Center before and after exercise. Exercise

was performed as a 30-min trotting on a treadmill. Trizol reagent (Invitrogen) was used to extract total RNA from leukocytes after exercise, according to the Invitrogen manual. In order to prevent contamination of genomic DNA, RNase-free DNase kit (Qiagen) was used according to the manufacturer's operating manual. Total RNA quantification was performed by using NanoDrop® ND-1000 Spectrophotometer. cDNAs were synthesized in a reaction with oligo-dT primers, moloney-murine leukemia virus (MMLV) reverse transcriptase (Promega), RNase inhibitor (Promega) and RNase-free ddH2O, which was incubated at 37℃ for 4 h. To confirm the *de novo* unique differentially expressed genes (DUDEGs) revealed by RNA-Seq, seven DUDEGs were analyzed by qRT-PCR amplification. The primers were designed using the PRIMER3 software (http://frodo.wi.mit.edu/primer3/) (Table 4.1).

The qRT-PCR conditions were as follows: an initial step of 94℃ for 10 min, 35 cycles of 94℃ for 30 sec, 60℃ for 30 sec, 72℃ for 30 sec, and final step of 72℃ for 10 min. PCR bands were normalized with glyceraldehyde-3-phosphate dehydrogenase (GAPDH) band. qRT-PCR products were visualized by gel electrophoresis on a 2.0% SeaKem LE agarose gel. cDNA was analyzed by BioRad CFX-96. All samples were measured in triplicate to ensure reproducibility, and Ct value was calculated using $2^{-\Delta\Delta Ct}$ method (Livak et al. 2001).

Table 4.1. qRT-PCR primer information such as the gene symbol, direction and sequence

Gene Symbol	Direction	Sequence
TIMELESS	Forward	TAG TGC CCT TGG GTA CTT GG
	Reverse	TGC TGG ATA AGG ATG GGA AG
EIF4A3	Forward	GCT GAT TTG ATT TGC CTT
	Reverse	GTT GTT GGG CAG GTC GTA GT
PIGW	Forward	GGG GCA GGA ATG TTC TAT CA
	Reverse	AAA GTC CAC AGC CAA AAT GG
ANK3	Forward	TGG CAG AAC GAG ACA TCA AG
	Reverse	ACA TGG CTT CCA TTT GCT TC
MSH3	Forward	AGC AGC AGA AAG ATG CCA TT
	Reverse	GCC TTT AAC GCT GCT GTT TC
SYNRG	Forward	TGT CTC AAC TCG GAC AGC AC
	Reverse	GAG TCA TCC AGG GTT CCT GA
ASGR2	Forward	ATC TGC GCA TCC TAG CTT GT
	Reverse	ATA TGA AAG GGG CTC GTG TG

4.3.3 Analysis of horse whole genome re-sequencing data

Whole genome re-sequencing data of Thoroughbred and Jeju domestic ponies

Whole-blood samples were collected from 18 Thoroughbred racing stallions of the Korean Racing Authority, and from four male and two female Jeju domestic ponies (*Equus caballus*) of the Jeju Provincial Livestock Institute, Korea. A 10 mL sample of blood was drawn from the carotid artery of each horse and was treated with heparin to prevent clotting. Genomic DNA was extracted and a quality check was carried out using fluorescence-based quantification on an agarose gel, a standard electrophoresis on a 0.6% agarose gel and, via a pulsed-field gel, using 200 ng of DNA. Manufacturers' instructions were followed to create a paired library of 500-bp fragments. This consisted of the following: purified genomic DNA fragments of less than 800 bp, fragments with blunt ends, fragments with 5′ phosphorylated ends, fragments with a 3′- dA overhang, some with adaptor-modified ends, purified ligation product, and a genomic DNA library. Following this, I generated sequence data using HiSeq 2000 (Illumina, Inc).

Reference genome assembly

Using the Burrows-Wheeler Aligner (Li et al. 2010) with the default setting, pair-end sequence reads were mapped to the reference horse genome (ftp://ftp.ensembl.org/pub/release-73/fasta/equus_caballus/dna/) (Table 4.5). The DNA re-sequencing data from this study have been submitted to the

NCBI Sequence Read Archive (SRA) database under the accession numbers SRA053569, SRA054885 and SRP017702.

Genotype calling and SNP calling

I used the following open-source software packages; Picard Tools, SAMtools, and the Genome analysis toolkit, for downstream processing and variant calling. Substitution calls were made with GATK UnifiedGenotyper20 and all calls with a Phred-scaled quality of less than 30 were filtered out. For each chromosome, I simultaneously inferred the phased haplotype and inputted the missing alleles for the entire set of Thoroughbred populations using BEAGLE (Browning et al. 2009).

Estimation of Nucleotide diversity, FST and Cross Population Extended Haplotype Homozygosity (XP-EHH) value

Nucleotide diversity and long run of homozygosity (LROH) of Thoroughbred and Jeju domestic ponies for each chromosome were calculated by VCFtools. Conventional F_{ST} (Wright 1949) and Reynolds F_{ST} (Reynolds et al. 1983) values were calculated for genes using Arlequin 3.5 (Excoffier et al. 2010) based on pairwise differences between the haplotypes of Thoroughbred and Jeju domestic ponies. In order to calculate F_{ST}, I used the horse genome to phase the haplotypes of the two populations. Also, to calculate F_{ST} by each gene region, I used the genomic information (Ensembl Genes71, EquCab2), namely the Ensembl reference annotated gene information. I selected the

genes of the top 1% of the empirical distribution (empirical p-value <0.01) (Teshima et al. 2006). The method Cross Population Extended Haplotype Homozygosity (XP-EHH) was used to detect selective sweeps using the software xpehh (Sabeti et al. 2007) (http://hgdp.uchicago.edu/Software/). For XP-EHH analysis, I used haplotype information for all SNPs of the entire autosome, and I calculated Extended Haplotype Homozygosity (EHH) and the log-ratio integrated EHH (iHH) for the pairwise test of the Thoroughbred and Jeju domestic pony populations. The log ratios were standardized to have a mean of 0 and a variance of 1, and p-values were assigned assuming a normal distribution. I selected SNPs with p-values < 0.01, which are considered to have strong selection signals. Then I apply a cutoff value of XP-EHH values < 0 for finding adaptation in the Thoroughbred. I chose genes related with these SNPs by identifying genes located within a 10kb (Sabeti et al. 2006) boundary of these SNPs. Since XP-EHH is not sensitive to allele frequencies, there is no need to stratify the data into frequency bins before determining significance. The p-values are empirical p-values; that is, a low p-value indicates that a locus is an outlier with respect to the rest of the genome. However, I note that loci detected as being under selection using this approach may be an under-representative sample of all truly selected loci; in particular, selection on standing variation and recessive loci are likely to be underrepresented (Teshima et al. 2006).

4.4 Results

4.4.1 Differences in the results of reference-based and *de novo*-based assembly and analysis

Transcriptome analysis results of reference-based analysis (RBA) and de novo-based analysis (DBA) showed a substantial difference in the number of transcript and differentially expressed genes (DEGs) identified. In RBA, for blood and skeletal muscle, 15,900 and 17,927 transcripts were found, respectively, among which 2,244 and 1,405 were unique transcripts. In DBA, the numbers of transcripts in skeletal muscle and blood were 18,057 and 19,413, respectively with 4,401 and 2,892 unique transcripts. The numbers transcripts in common between RBA and DBA were 13,656 for skeletal muscle tissue and 16,521 for blood (Figure 4.1a and Figure 4.2a). When the sample variance of RBA and DBA in skeletal muscle and blood were compared using multidimensional scaling (MDS) plot, the results for the two analyses were almost identical. The skeletal muscle samples were clustered into two subgroups: before and after exercise, but the blood samples did not show any clustering (Figure 4.1b, Figure 4.2b and Figure 4.3). In RBA, the number of DEG in skeletal muscle and blood were 2,818 and 455, respectively with 2,200 and 427 DEGs being unique to RBA. In DBA, the number of DEG in skeletal muscle and blood were 1,652 and 595, respectively with 1,034 and 567 unique DEGs. The number of DEGs identified by both RBA and DBA were 618 and 28 in skeletal muscle and blood, respectively (Figure 4.1c and Figure 4.2c). These DEGs were

compared using Heatmap visualization to examine their expression pattern in each analysis. The expression pattern was similar, however, the intensity of the expression was higher with DBA (Figure 4.1d and Figure 4.2d). Overall, in comparison to RBA, DBA identified a higher number of transcripts but a lower number of DEGs. I detected SNPs from two different next-generation sequencing methods (WGS and RNA-seq) and two different assembly methods (RBA and DBA) for each Thoroughbred sample (F1, F2 and F3 = male, S3 = female). The number and rate of SNPs in DBA of RNA-seq was 108,158 (0.031%), 110,502 (0.031%), 105,920 (0.03%) and 101,887 (0.029%) in F1, F2, F3 and S3 respectively, and the number and rate of SNPs in from RBA of RNA-seq were 284,859 (0.012%), 287,286 (0.012%), 276,241 (0.011%) and 265,729 (0.011%) in F1, F2, F3 and S3, respectively (Table 4.2).

Table 4.2. The number and rate of SNPs from different next-generation sequencing method (DNA and RNA sequencing) and different reference genome assembly in each Thoroughbred horse sample (F1, F2 and F3 = male, S3 = female).

SNP detection		F1	F2	F3	S3
Sequence	Reference				
WGS	DNA	3,797,464 (0.153%)	3,716,018 (0.150%)	3,759,757 (0.152%)	3,628,882 (0.147%)
	cDNA	35,591 (0.08%)	38,275 (0.082%)	36,058 (0.077%)	36,121 (0.078%)
RNA-seq	DNA	284,859 (0.012%)	287,286 (0.012%)	276,241 (0.011%)	265,729 (0.011)
	cDNA	29,507 (0.063%)	30,570 (0.066%)	30,145 (0.065%)	29,518 (0.063%)
RNA-seq	Trinity(*de novo*)	108,158 (0.031%)	110,502 (0.031%)	105,920 (0.030%)	101,887 (0.029%)

Table 4.3. List of basic stats such as the number of transcripts, components, and contigs N50 value in RNA-seq whole reads and unmapped reads by trinity de-novo assembly.

SAMPLE ID	READS	TRANSCRIPTS	COMPONENTS	CONTIG N50
BF1B	Whole	215506	159877	1814
	Unmapped	5597	4914	421
BF1P	Whole	200291	169168	1358
	Unmapped	775	717	373
BF2B	Whole	219397	186218	1114
	Unmapped	6197	5545	388
BF2P	Whole	201725	167243	1418
	Unmapped	634	603	370
BF3B	Whole	194689	160340	1584
	Unmapped	1070	999	351
BF3P	Whole	197659	165630	1389
	Unmapped	1111	1034	354
BS1B	Whole	206071	173251	1294
	Unmapped	5277	4624	431
BS1P	Whole	221676	184356	1375
	Unmapped	6285	5478	454
BS2B	Whole	213508	174922	2051
	Unmapped	3210	2881	411
BS2P	Whole	235880	193273	1106
	Unmapped	697	663	352
BS3B	Whole	189580	155230	1673
	Unmapped	623	577	374
BS3P	Whole	176267	146704	1430
	Unmapped	5440	4772	458
MF1B	Whole	108250	95928	826
	Unmapped	241	259	361

MF1P	Whole	96293	86030	833
	Unmapped	246	237	342
MF2B	Whole	84468	75565	813
	Unmapped	4478	4136	384
MF2P	Whole	86727	78188	725
	Unmapped	432	414	376
MF3B	Whole	98665	88042	964
	Unmapped	621	599	372
MF3P	Whole	74551	68106	674
	Unmapped	2403	2268	354
MS1B	Whole	100646	89752	774
	Unmapped	165	158	413
MS1P	Whole	93819	84093	858
	Unmapped	164	160	369
MS2B	Whole	85543	76894	977
	Unmapped	682	661	369
MS2P	Whole	92927	82949	992
	Unmapped	710	688	362
MS3B	Whole	76011	68916	862
	Unmapped	191	182	401
MS3P	Whole	81533	73773	841
	Unmapped	131	125	480

Table 4.4. Number of annotated transcripts from RNA-seq unmapped reads by trinity de-novo assembly. The number in the parentheses is the number of transcripts that were not included in the results of the reference-based analysis.

Sample ID	**BF1B**	**BF1P**	**BF2B**	**BF2P**	**BF3B**	**BF3P**
Annotated transcripts	53(3)	5(1)	55(4)	7(0)	12(1)	12(1)
Sample ID	**BS1B**	**BS1P**	**BS2B**	**BS2P**	**BS3B**	**BS3P**
Annotated transcripts	49(3)	83(5)	44(3)	14(0)	10(1)	118(5)
Sample ID	**MF1B**	**MF1P**	**MF2B**	**MF2P**	**MF3B**	**MF3P**
Annotated transcripts	8(1)	9(2)	147(21)	14(3)	18(4)	59(8)
Sample ID	**MS1B**	**MS1P**	**MS2B**	**MS2P**	**MS3B**	**MS3P**
Annotated transcripts	6(2)	6(1)	25(6)	28(15)	14(4)	6(1)

Table 4.5. Basic information of 4 horses re-sequencing data

Individual	HORSE1(F2)	HORSE2(F1)	HORSE3(F3)	HORSE4(S3)
Gender	Male	Male	Male	Female
Sample	Blood	Blood	Blood	Blood
Location	BGI	BGI	BGI	BGI
Sequencing depth	10X	10X	10X	10X
Read Length	90	90	90	90
Encoding	illumina 1.5	illumina 1.5	illumina 1.5	illumina 1.5
No. Raw Read	314,274,368	302,024,230	296,743,324	282,784,938
No. Read	285,815,336	273,659,950	270,715,242	256,521,664
Trimming	5bp	5bp	5bp	5bp
No. mapped read	261,682,398	248,053,944	247,267,550	235,131,946
Mapping rate (%)	91.56	90.64	91.34	91.66

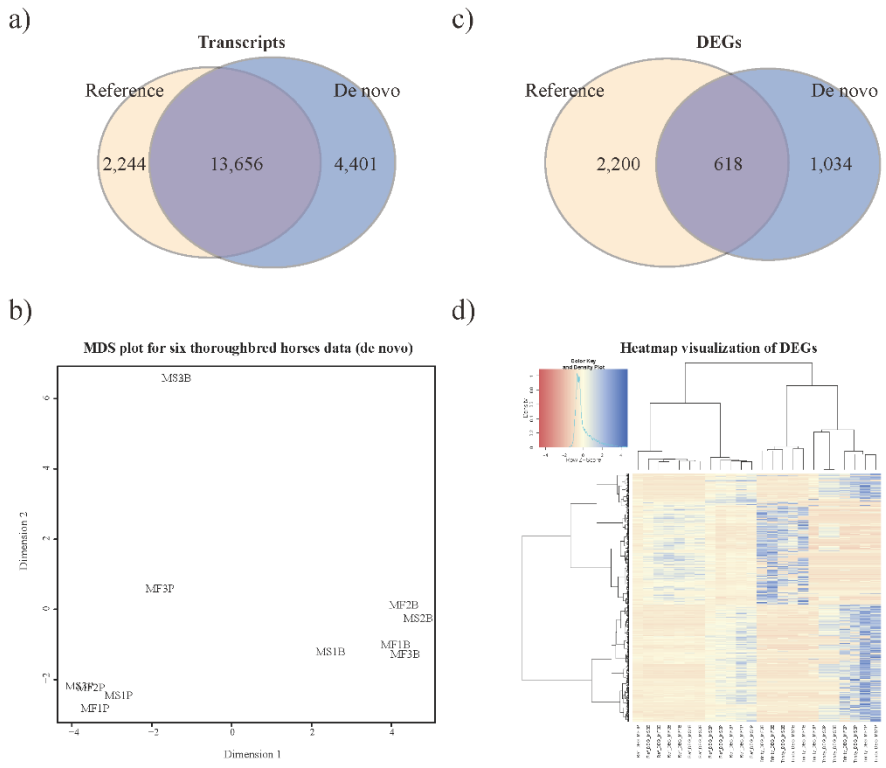


Figure 4.1. Summary of comparative analysis between de novo assembly and reference genome assembly from RNA-seq data of skeletal muscle from six Thoroughbreds before and after exercise (Total 12 samples).

a) The number of transcripts in common between de novo assembly and reference genome assembly b) MDS plot of six Thoroughbreds before and after exercise using de novo assembly. c) The number of DEGs between de novo assembly and reference genome assembly. d) Heat-map visualization of common DEGs between de novo assembly and reference genome assembly: rows represent DEGs from skeletal muscle and columns represent assemble method from 6 horse samples (*First ‘B’ is for blood and ‘M’ is for muscle. ‘F1’, ‘F2’, ‘F3’ and ‘S3’ are horse samples. Last ‘B’ is for ‘before exercise’ and ‘P’ is for ‘after exercise’)

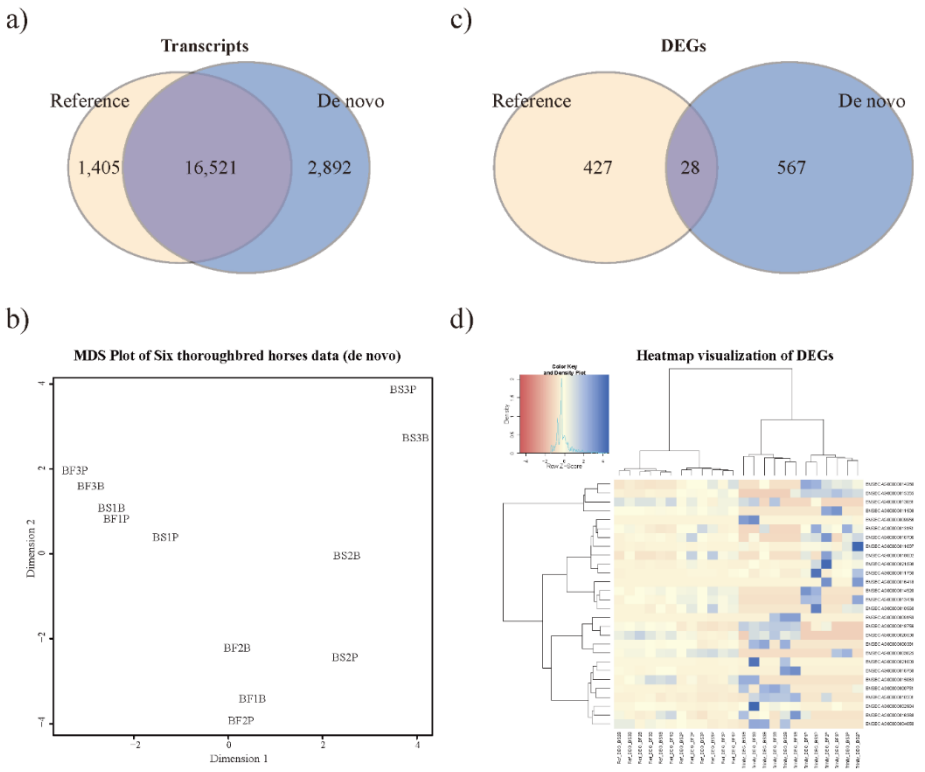


Figure 4.2. Summary of comparative analysis between *de novo* assemble and reference genome assemble from blood in six Thoroughbred horses before and after exercise RNA-seq data (Total 12 samples).

a) The number of common transcripts of 12 samples between de novo assemble and reference genome assemble b) MDS plot of six Thoroughbred horses before and after exercise using de novo assemble. c) The number of DEGs between de novo assemble and reference genome assemble. d) Heat-map visualization of common DEGs between de novo assemble and reference genome assemble: rows represent DEGs from blood and columns represent assemble method from 6 horse samples (*First ‘B’ is for Blood and ‘M’ is for muscle. ‘F1’, ‘F2’, ‘F3’ and ‘S3’ are horse samples. Last ‘B’ is for ‘before exercise’ and ‘P’ is for ‘after exercise’)

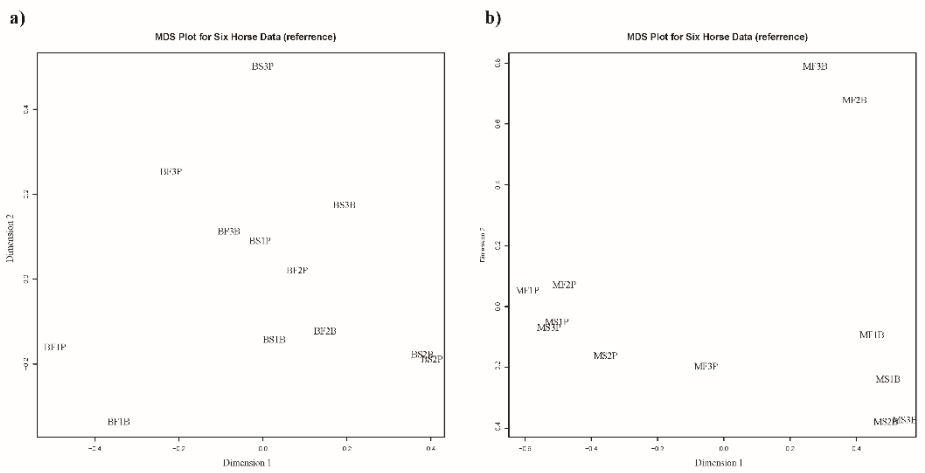


Figure 4.3. MDS plot of six Thoroughbred horses before and after exercise using reference genome assemble in RNA-seq.

a) MDS plot of blood tissue in six Thoroughbred horse before and after exercise. b) MDS plot of skeletal muscle tissue in six Thoroughbred horse before and after exercise. (*First ‘B’ is for Blood and ‘M’ is for muscle. ‘F1’, ‘F2’, ‘F3’ and ‘S3’ are horse samples. Last ‘B’ is for ‘before exercise’ and ‘P’ is for ‘after exercise’)

4.4.2 Identification of de novo unique differentially expressed genes (DUDEGs) before and after exercise

I identified DUDEGs from RNA-seq data using the expression profiles of genes in skeletal muscle and blood samples taken from six Thoroughbreds before and after exercise. There were a total of 1,034 significant DUDEGs (519 up-regulated, 515 down-regulated) in skeletal muscle and 567 (314 up-regulated, 253 down-regulated) in blood (FDR <0.01). Among them, 456 (61 up-regulated, 395 down-regulated) in skeletal muscle and 205 (93 up-regulated, 112 down-regulated) in blood were annotated.

4.4.3 Validation of DUDEGs in horse RNA-seq data using RT-PCR

I performed qRT-PCR to validate the DUDEGs detected in horse blood. The seven genes (*TIMELESS, EIF4A3, PGIW, ANK3, MSH3, SYNRG, ASGR2*: 2 up-regulated and 5 down-regulated) were randomly selected with conceptually new DEGs and logFC > 2 in blood (Table 4.8 and Table 4.9). The expression levels of DUDEGs between RNAseq and qRT-PCR were highly similar (Figure 4.4). The results confirmed that DUDEGs identified in this study were reliable.

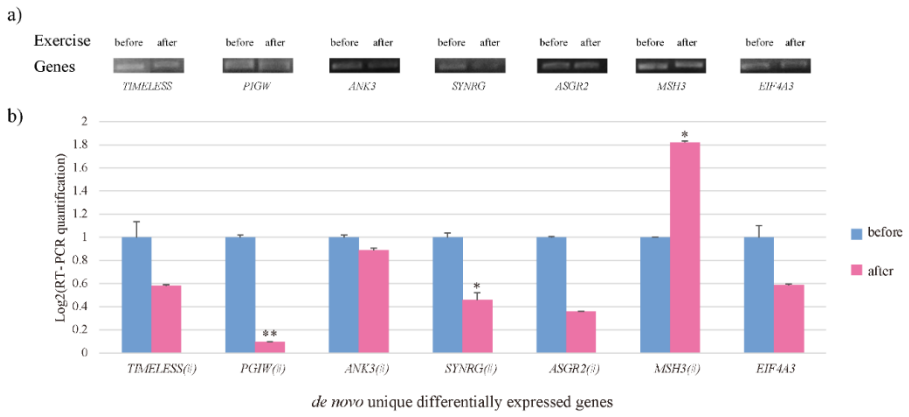


Figure 4.4. qRT-PCR validation of de novo unique differentially expressed genes (DUDEGs) identified from the RNA-seq data set of Thoroughbred horses before and after exercise

a) qRT-PCR of six DUDEG in horses before exercise and after exercise. b) qRT-PCR results depicted as Ct value was calculated using $2^{-\Delta\Delta Ct}$ method. *: p-value < 0.05. **: p-value < 0.01. #: The expression patters of genes supported the result of my analysis

4.4.4 Functional annotation of DUDEGs

I summarized the highest biological process gene ontology (BP-GO) of DUDEGs in skeletal muscle sample taken before and after exercise from six Thoroughbred RNA-seq data (Figure 4.5). The other BP-GO of DUDEGs were summarized separately (Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9). The most significantly enriched terms among up-regulated genes in skeletal muscle were 'biological adhesion', 'biological regulation', 'death', 'growth', 'immune system process', 'locomotion', 'multi-organism process', and 'response to stimulus' (P-value = 6.29E-02, P-value = 9.57E-04, P-value = 1.75E-06 , P-value = 7.56E-03 , P-value = 2.23E-13 , P-value = 3.48E-03 , P-value = 5.71E-04, and P-value = 5.16E-08, respectively). While, the most significantly enriched terms among down-regulated genes in skeletal muscle were 'cellular component organization', 'cellular process', 'establishment of localization', 'localization', and 'metabolic process' (P-value = 4.94E-04, P-value = 9.12E-04, P-value = 7.09E-02, P-value = 6.87E-02, P-value = 5.28E-02, and P-value = 1.89E-2, respectively). 'Developmental process' was the most significantly enriched term in skeletal muscle among both up and down-regulated genes. However, no terms were highlighted as being significantly enriched in blood. I summarized the cellular components and molecular function gene ontology of DUDEGs in RNA-seq data from skeletal muscle and blood of six Thoroughbreds before and after exercise. Enriched KEGG pathways analysis using DUDEGs revealed that up-regulated genes in skeletal muscle and blood are associated with exercise-induced stress. The most significantly enriched terms in skeletal muscle were 'p53 signaling pathway', 'regulation of actin cytoskeleton', 'JAK-STAT signaling pathway', 'MAPK

signaling pathway', 'cell adhesion molecules', 'cytokine-cytokine receptor interaction', 'bladder cancer', and 'pathways in cancer'. In addition, the two terms 'colorectal cancer' and 'biosynthesis of unsaturated fatty acids' were significantly enriched in blood (Table 4.6).

Table 4.6. Enriched KEGG pathways associated with DEGs in two tissue such as skeletal muscle and blood. For each set of up-regulated and down-regulated. DEG in skeletal muscle and blood, a KEGG pathway enrichment analysis was performed. Starting from the right, the table shows: tissue type, status of regulation, KEGG pathway terms, higher-level KEGG pathway terms, and the highest level of KEGG pathway terms.

Highest KEGG	Higher KEGG	KEGG	Blood		Muscle	
			UP	DOWN	UP	DOWN
Cellular Processes	Cell growth and death	Cell cycle				V
		p53 signaling pathway			V	
Environmental Information Processing	Cell motility	Regulation of actin cytoskeleton			V	
	Membrane transport	ABC transporters				V
	Signal transduction	Jak-STAT signaling pathway			V	
		MAPK signaling pathway			V	
		Notch signaling pathway				V
		Phosphatidylinositol signaling system				V
	Signaling molecules and interaction	Cell adhesion molecules (CAMs)			V	
		Cytokine-cytokine receptor interaction			V	

Genetic Information Processing	Folding, sorting and degradation	Ubiquitin mediated proteolysis				V
	Replication and repair	Non-homologous end-joining				V
Human Diseases	Cancers	Colorectal cancer	V			
		Bladder cancer		V	V	
		Pathways in cancer			V	
		Small cell lung cancer		V		
Metabolism	Carbohydrate metabolism	Butanoate metabolism				V
		Inositol phosphate metabolism				V
	Glycan biosynthesis and metabolism	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis		V		V
	Lipid metabolism	Biosynthesis of unsaturated fatty acids	V			
Organismal Systems	Immune system	B cell receptor signaling pathway			V	
		Hematopoietic cell lineage			V	
		Natural killer cell mediated cytotoxicity			V	
		Toll-like receptor signaling pathway			V	
		T cell receptor signaling pathway			V	

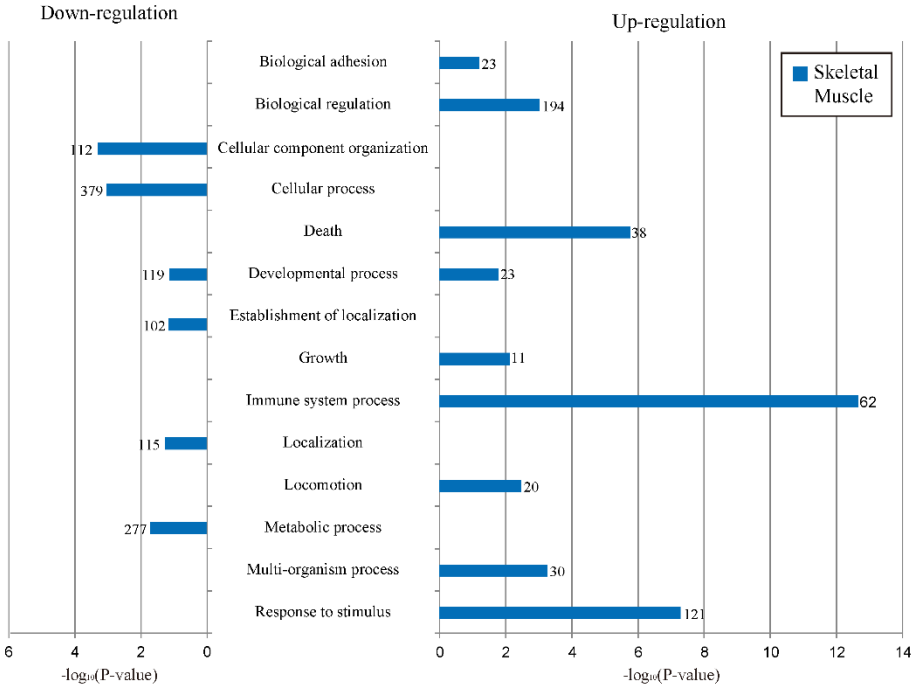


Figure 4.5. Biological process GO terms of tissues specific DEGs between before and after exercise in Thoroughbred.

Up–regulated genes indicate higher activation after exercise than before and down-regulation genes indicate lower activation after exercise than before exercise.

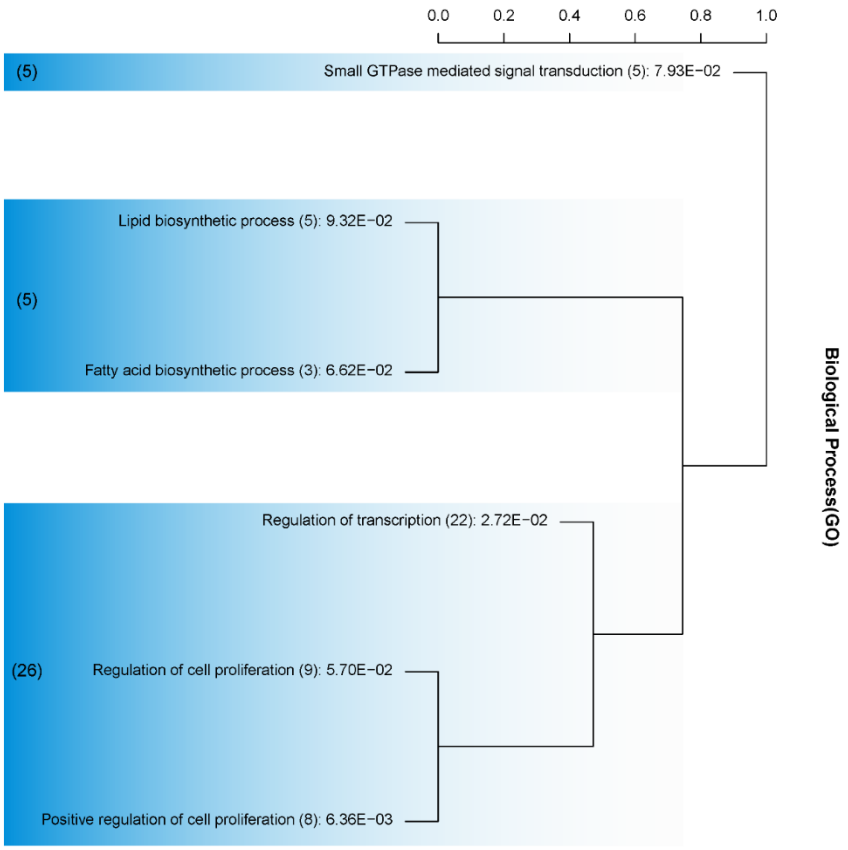


Figure 4.6. Hierarchical clustering of biological process GO terms associated with up-regulated DEGs in blood.

The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block.

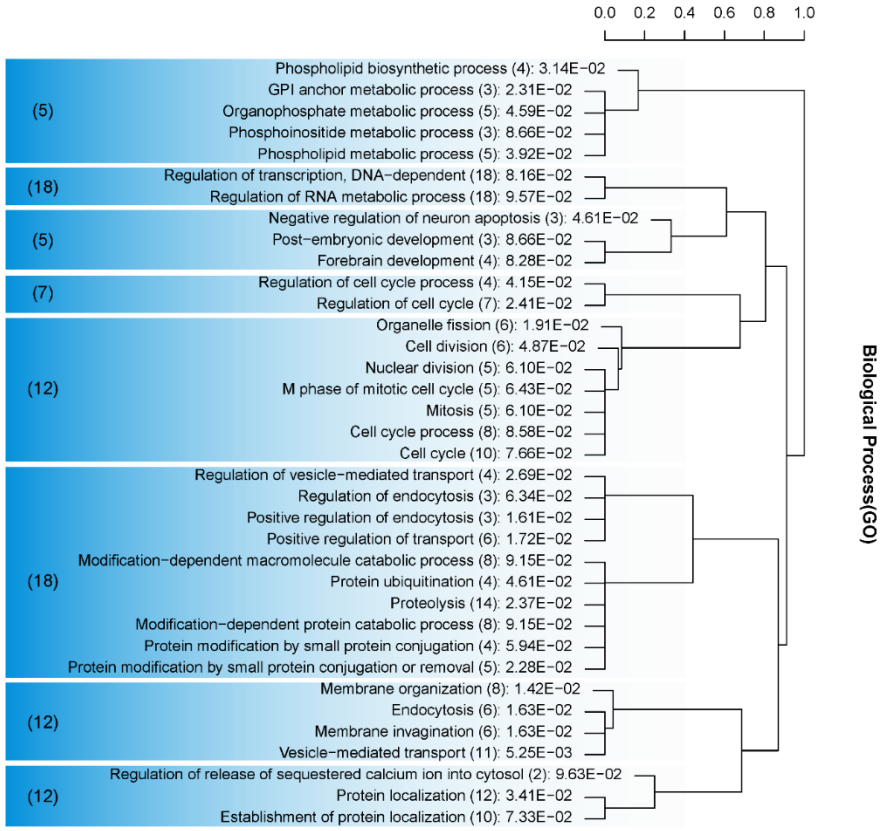


Figure 4.7. Hierarchical clustering of biological process GO terms associated with down-regulated DEGs in blood.

The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block.

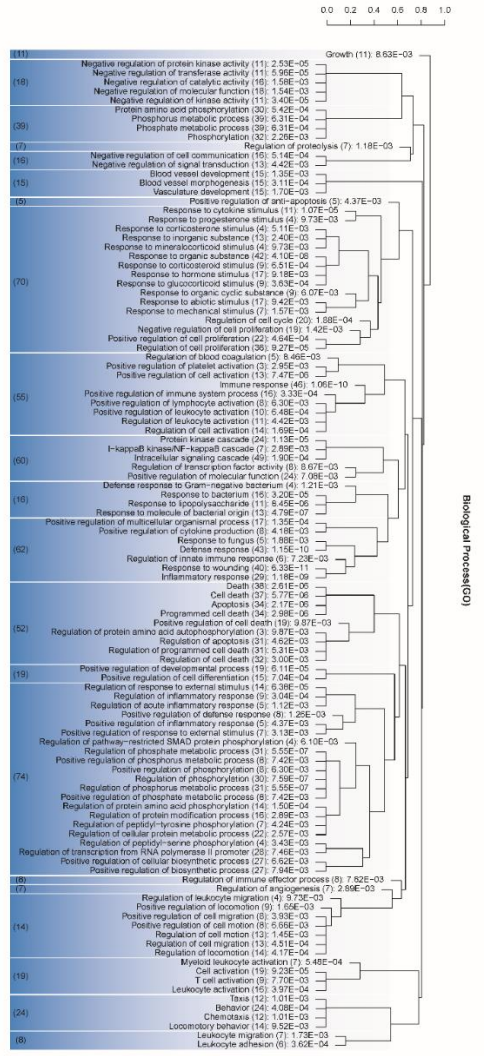


Figure 4.8. Hierarchical clustering of biological process GO terms associated with up-regulated DEGs in muscle.

The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block.

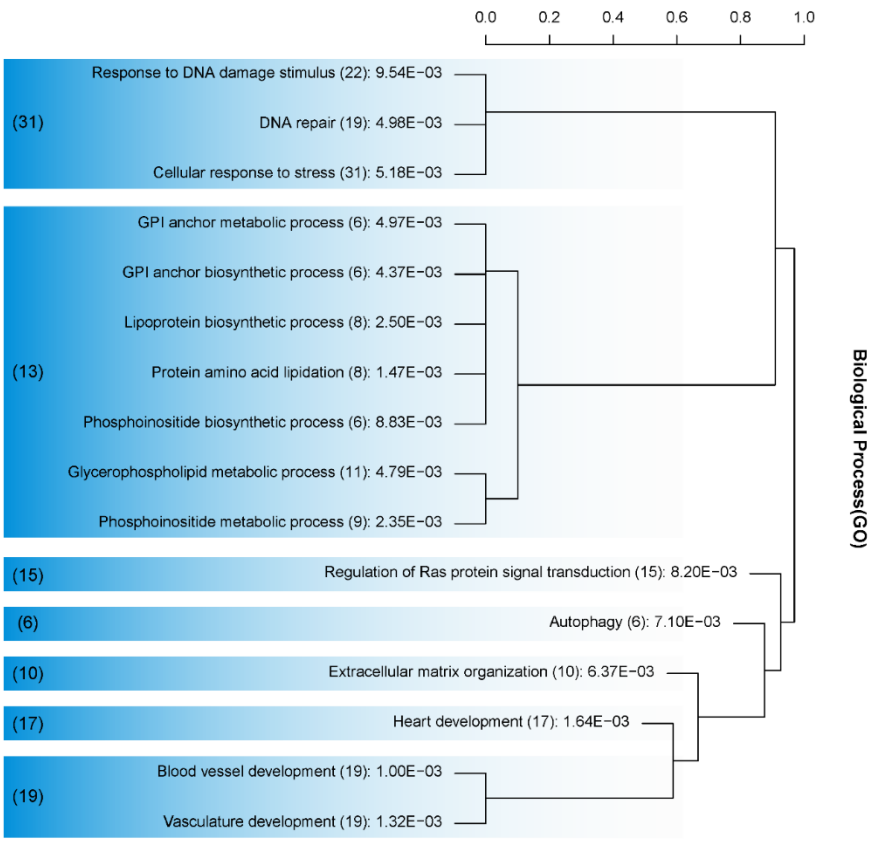


Figure 4.9. Hierarchical clustering of biological process GO terms associated with down-regulated DEGs in muscle.

The gene list of each GO term clustered using DAVID was compared to calculate the distance between the GO terms. For a distance value >0.5, GO terms were re-clustered, and GO term groups are shown as light-blue graduated blocks. The number of genes associated with the re-clustered GO term group is shown on the left side of the block. .

4.4.5 Integration of DUDEG and selected gene associated with nucleotide diversity, F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH)

The genes under selection were further investigated using nucleotide diversity, F_{ST} and XP-EHH from WGS of Thoroughbred and Jeju pony (skeletal muscle = 1,033 and blood = 567). I found 12 genes (*ZNF592, CD58, C1orf162, USP37, FOXM1, TIMELESS, TRMT1, CALR, ASNA1, EIF4A3, SYNRG* and *FADS1*) in blood and 14 genes (*HERC2, CHD9, DDX28, CAPZA1, TSEN15, CHMP4C, FOXO3, PLD2, ANKRD13D, UNKL, CBFA2T2, NECAB3, SLC25A29* and *FBLN1*) in skeletal muscle that were both identified as DUDEGs and implicated in F_{ST} analysis as a selected gene (Table 4.8). The F_{ST} distribution histogram for the pair of horse breeds is shown in Figure 4.11. Between DUDEGs and XP-EHH (selected genes in the Thoroughbred), I found 11 (*ANK3, ZNF592, TOR1AIP1, TIMELESS, INSR, MED15, ZNF567, EIF4A3, EXOC6B, PPP4R2* and *DYRK1A*) and 48 common genes (*URB2, NEO1, CORO2B, CPE, FUK, SYCE1L, TEC, RELL1, ZNF775, RGS4, FAM46C, PTGFRN, DENND2C, CTTNBP2NL, VAV1, BARX2, RAB31, VPS4B, CHMP4C, DEPTOR, ATG5, FOXO3, OSBPL7, PEX12, PIK3R5, PITX1, MARCH3, LMNB1, ST8SIA4, RASGRF2, ARSB, INPP4A, RNF144A, PPP4R2, FRMD4B, SPATA13, SLC7A1, CAB39L, B3GALT1, MECOM, PARP14, NPR3, TGM3, DHX35, AUH, C14orf102, COL27A1* and *HLCS*) in blood and skeletal muscle, respectively (Table 4.9). Among them, three genes, *TIMELESS, EIF4A3* and *ZNF592*, in blood, and two genes, *CHMP4C* and *FOXO3*, in skeletal muscle were shown to be significant in all three analyses (DUDEG, F_{ST}, and XP-EHH). In comparison to Jeju pony, the Thoroughbred

showed relatively low levels of nucleotide diversity at three out of the five identified genes (Table 4.7 and Figure 4.10).

Table 4.7. co-matching genes between the DEGs, selected genes associated with F_{ST} (F_{ST} cut-off value top 5% with empirical p-value < 0.05) and Thoroughbred selected genes associated with XP-EHH (XP-EHH cut-off value empirical p-value < 0.01 and XP-EHH value < -3.51551 significant SNPs)

Sample Tissue	Ens ID	CHR	Start	End	DEG logFC	DEG P-value	DEG FDR	Gene symbol	Reynolds	F_{ST}	SNP region	XP-EHH	XP-EHH P-value
Blood	ENSECAG00000015925	11	2814986	2828951	-17.23	1.81E-34	2.94E-31	*EIF4A3* (eukaryotic translation initiation factor 4A3)	0.83481	0.56604	2821795	-3.52526	9.73E-03
	ENSECAG00000016283	1	92397477	92416004	-14.5988	2.23E-26	5.06E-24	*ZNF592* (zinc finger protein 592)	0.67059	0.48859	92387687	-3.55925	8.83E-03
											92397597	-3.78677	4.48E-03
											92398170	-3.74169	5.15E-03
											92398472	-3.78551	4.50E-03
											92398776	-3.83657	3.84E-03
											92398791	-4.18477	1.23E-03
											92398828	-3.56243	8.74E-03
											92398986	-3.5556	8.92E-03
											92399035	-3.76415	4.81E-03
											92402102	-3.60557	7.72E-03
											92402157	-3.58028	8.31E-03
											92402965	-3.58305	8.24E-03

											92403116	-3.61357	7.54E-03
											92403392	-3.55258	9.00E-03
	ENSECAG00000002892	6	74016170	74027809	9.80616	1.34E-12	6.75E-11	*TIMELESS* (timeless circadian clock)	0.67349	0.49007	74019099	-4.02647	2.09E-03
											74019114	-4.05026	1.93E-03
											74019135	-4.28442	8.68E-04
											74019141	-4.56193	3.15E-04
											74019149	-4.42328	5.27E-04
											74019152	-4.43649	5.02E-04
											74019153	-4.81799	1.16E-04
											74019241	-4.96438	6.40E-05
Muscle	ENSECAG00000022647	9	6229159	6251198	12.97629	5.96E-22	7.12E-20	*CHMP4C* (charged multivesicular body protein 4C)	1.29183	0.72523	6257522	-3.56699	8.63E-03
	ENSECAG00000024499	10	58506453	58523225	1.941151	0.001093	0.008744	*FOXO3* (forkhead box O3)	0.91477	0.59939	58521033	-3.5866	8.15E-03
											58521300	-3.67329	6.32E-03

Table 4.8. Common genes between DEGs and selected genes associated with F_{ST} (F_{ST} cut-off value top 5% with empirical p-value < 0.05)

Sample Tissue	Ens ID	CHR	Start	End	DEG logFC	DEG P-value	DEG FDR	Gene symbol	Reynolds	Fst
Blood	ENSECAG00000016283	1	92397477	92416004	-14.5988	2.23E-26	5.06E-24	ZNF592 (zinc finger protein 592)	0.67059	0.48859
	ENSECAG00000024391	5	52474368	52488930	-12.7633	6.63E-21	5.66E-19	CD58 (CD58 molecule)	0.63486	0.46999
	ENSECAG00000018585	5	56885158	56885977	13.73641	5.74E-24	7.51E-22	C1orf162 (chromosome 1 open reading frame 162)	0.66702	0.48676
	ENSECAG00000018200	6	8097875	8184891	2.551938	3.11E-04	7.70E-03	USP37 (ubiquitin specific peptidase 37)	0.68323	0.49502
	ENSECAG00000019129	6	30817902	30826537	-2.66297	6.60E-05	1.84E-03	FOXM1 (forkhead box M1)	1.04819	0.64943
	ENSECAG00000002892	6	74016170	74027809	9.80616	1.34E-12	6.75E-11	TIMELESS (timeless circadian clock)	0.67349	0.49007
	ENSECAG00000016149	7	45346016	45351774	10.48754	1.60E-14	9.18E-13	TRMT1 (tRNA methyltransferase 1 homolog)	1.11214	0.67114
	ENSECAG00000008164	7	45483000	45486353	-16.153	4.48E-31	2.71E-28	CALR (calreticulin)	0.76759	0.53587
	ENSECAG00000007103	7	45610152	45615940	12.03488	6.08E-19	4.44E-17	ASNA1 (arsA arsenite transporter, ATP-binding, homolog 1)	0.82578	0.56211
	ENSECAG00000015925	11	2814986	2828951	-17.23	1.81E-34	2.94E-31	EIF4A3 (eukaryotic translation initiation factor 4A3)	0.83481	0.56604
	ENSECAG00000008489	11	36666971	36750948	13.48276	3.21E-23	3.65E-21	SYNRG (synergin, gamma)	0.69484	0.50084
	ENSECAG00000023280	12	21722550	21731721	-11.6068	3.32E-18	2.27E-16	FADS1 (fatty acid desaturase 1)	0.6729	0.48977
Muscle	ENSECAG00000017677	1	1.14E+08	1.14E+08	13.61197	1.19E-23	2.08E-21	HERC2 (HECT and RLD domain containing E3 ubiquitin protein ligase 2)	0.6413	0.47339
	ENSECAG00000011505	3	5744217	5890102	17.52678	1.84E-35	2.96E-32	CHD9 (chromodomain helicase DNA binding protein 9)	0.7025	0.50466
	ENSECAG00000006451	3	18334353	18335975	9.902519	5.43E-13	1.49E-11	DDX28 (DEAD (Asp-Glu-Ala-Asp) box polypeptide 28)	0.7841	0.54347
	ENSECAG00000016995	5	55853345	55897158	15.38217	1.21E-28	6.53E-26	CAPZA1 (capping protein (actin filament) muscle Z-line, alpha 1)	0.92616	0.60393

ENSECAG00000003249	6	35501724	35501933	3.254749	7.49E-06	1.06E-04	TSEN15 (TSEN15 tRNA splicing endonuclease subunit)	0.72986	0.51802
ENSECAG00000022647	9	6229159	6251198	12.97629	5.96E-22	7.12E-20	CHMP4C (charged multivesicular body protein 4C)	1.29183	0.72523
ENSECAG00000024499	10	58506453	58523225	1.941151	1.09E-03	8.74E-03	FOXO3 (forkhead box O3)	0.91477	0.59939
ENSECAG00000014508	11	49675720	49688013	9.17594	6.05E-11	1.33E-09	PLD2 (phospholipase D2)	0.66424	0.48534
ENSECAG00000007911	12	27072526	27083029	10.73278	5.26E-15	1.91E-13	ANKRD13D (ankyrin repeat domain 13 family, member D)	0.66014	0.48322
ENSECAG00000010641	13	41334685	41377271	9.883918	4.35E-13	1.23E-11	UNKL (unkempt family zinc finger-like)	0.66723	0.48687
ENSECAG00000025052	22	24529033	24564706	10.39155	4.33E-14	1.38E-12	CBFA2T2 (core-binding factor, runt domain, alpha subunit 2; translocated to, 2)	0.79744	0.54952
ENSECAG00000012086	22	24576179	24592061	2.743281	7.61E-05	8.68E-04	NECAB3 (N-terminal EF-hand calcium binding protein 3)	0.72823	0.51724
ENSECAG00000006302	24	42214122	42226036	2.730814	3.77E-05	4.67E-04	SLC25A29 (solute carrier family, member 29)	0.7676	0.53587
ENSECAG00000018101	28	41466402	41532567	12.14179	3.38E-19	2.45E-17	FBLN1 (fibulin 1)	0.71941	0.51296

Table 4.9. Common genes between DEGs and selected genes associated with XP-EHH : XP-EHH cut-off value empirical p-value < 0.01 and XP-EHH value < -3.51551 significant SNPs in Thoroughbred were selected and > 1.73481 significant SNPs in Jeju domestic pony were selected

Sample Tissue	Ensembl ID	CHR	Start	End	DEG logFC	DEG P-value	DEG FDR	Gene symbol	# SNP	Mean of XP-WHH	Mean of XP-EHH P-value
Blood	ENSECAG00000009700	1	49551464	49778390	12.20373	1.89E-19	1.46E-17	ANK3 (ankyrin 3)	387	-3.86713	0.004582
	ENSECAG00000016283	1	92397477	92416004	-14.5988	2.23E-26	5.06E-24	ZNF592 (zinc finger protein 592)	14	-3.6937	0.006521
	ENSECAG00000000872	5	17363920	17397208	-14.2221	1.38E-25	2.57E-23	TOR1AIP1 (torsin A interacting protein 1)	5	-4.12532	0.00158
	ENSECAG00000002892	6	74016170	74027809	9.80616	1.34E-12	6.75E-11	TIMELESS (timeless circadian clock)	8	-4.44565	0.000802
	ENSECAG00000010127	7	4751960	4988487	12.37929	6.12E-20	4.99E-18	INSR (insulin receptor)	5	-3.80987	0.004652
	ENSECAG00000016841	8	178527	243772	-13.9264	2.75E-24	3.77E-22	MED15 (mediator complex subunit 15)	39	-3.6645	0.006726
	ENSECAG00000017599	10	7457036	7478256	16.37556	9.76E-32	6.44E-29	ZNF567 (zinc finger protein 567)	46	-3.6854	0.006615
	ENSECAG00000015925	11	2814986	2828951	-17.23	1.81E-34	2.94E-31	EIF4A3 (eukaryotic translation initiation factor 4A3)	1	-3.52526	0.009726
	ENSECAG00000019424	15	30121373	30696569	13.13288	3.3E-22	3.3E-20	EXOC6B (exocyst complex component 6B)	9	-3.94019	0.003525
	ENSECAG00000000689	16	17404324	17450056	-12.9826	8.86E-22	8.53E-20	PPP4R2 (protein phosphatase 4, regulatory subunit 2)	23	-3.89151	0.003803
	ENSECAG00000024965	26	33662424	33750590	-15.2732	8.89E-29	2.81E-26	DYRK1A (dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A)	7	-3.63808	0.007152
Muscle	ENSECAG00000012338	1	68213418	68243651	-2.78093	9.74E-05	0.001084	URB2 (URB2 ribosome biogenesis 2 homolog)	46	-3.70844	0.006394
	ENSECAG00000024743	1	1.2E+08	1.2E+08	2.216988	0.000535	0.004777	NEO1 (neogenin 1)	9	-3.81447	0.00455

ENSECAG00000014020	1	1.24E+08	1.24E+08	3.265104	3.08E-06	4.58E-05	CORO2B (coronin, actin binding protein, 2B)	37	-3.66469	0.006936
ENSECAG00000019874	2	19784715	19831890	9.284013	3.77E-11	8.48E-10	INPP5B (inositol polyphosphate-5-phosphatase, 75kDa)	3	1.760273	0.009301
ENSECAG00000020498	2	46247154	46251383	3.038198	6.32E-06	9E-05	C1orf174 (chromosome 1 open reading frame 174)	10	1.825307	0.007927
ENSECAG00000009256	2	69289564	69318502	2.829719	0.000587	0.005185	CPE (carboxypeptidase E)	11	-3.75014	0.00526
ENSECAG00000000244	3	2577934	2838642	3.18674	3.53E-06	5.19E-05	ZNF423 (zinc finger protein 423)	3	1.746867	0.009664
ENSECAG00000012351	3	23181694	23195306	2.432645	0.000194	0.001993	FUK (fucokinase)	2	-3.54777	0.009122
ENSECAG00000026996	3	25675996	25681658	8.656818	1.14E-09	2.24E-08	SYCE1L (synaptonemal complex central element protein 1 like)	43	-3.71331	0.006133
ENSECAG00000024389	3	80719113	80802247	-2.49741	0.000297	0.002887	TEC (tec protein tyrosine kinase)	11	-3.66016	0.006871
ENSECAG00000016126	3	89324004	89367130	-2.84043	7.72E-05	0.000878	RELL1 (RELT-like 1)	22	-4.35804	0.003112
ENSECAG00000016405	4	1.02E+08	1.02E+08	14.80328	1.98E-27	8.79E-25	ZNF775 (zinc finger protein 775)	6	-3.61461	0.007576
ENSECAG00000017203	5	74951	78919	-2.60874	0.000161	0.001709	BTG2 (BTG family, member 2)	3	1.796523	0.00838
ENSECAG00000018261	5	33996082	34163712	-3.01813	1.43E-05	0.000195	RGS4 (regulator of G-protein signaling 4)	3	-3.68571	0.006294
ENSECAG00000019626	5	51494441	51512333	-2.16498	0.001265	0.009912	FAM46C (family with sequence similarity 46, member C)	5	-3.54222	0.009277
ENSECAG00000017101	5	52109720	52148091	10.72769	4.91E-15	1.79E-13	PTGFRN (prostaglandin F2 receptor inhibitor)	190	-3.83333	0.004681
ENSECAG00000019594	5	54125841	54157241	-2.43055	0.000744	0.006297	DENND2C (DENN/MADD domain containing 2C)	1	-3.70886	0.005683
ENSECAG00000010817	5	56038131	56074031	-3.01296	2.79E-05	0.000356	CTTNBP2NL (CTTNBP2 N-terminal like)	63	-3.81149	0.005189
ENSECAG00000014896	7	4329680	4385853	-10.3104	9.78E-14	2.96E-12	VAV1 (vav 1 guanine nucleotide exchange factor)	30	-3.81222	0.005233
ENSECAG00000000616	7	38020741	38088004	10.37942	4.58E-14	1.46E-12	BARX2 (BARX homeobox 2)	21	-3.69688	0.006732
ENSECAG00000006886	8	34550671	34667158	-2.28078	0.001102	0.008807	RAB31 (RAB31, member RAS oncogene family)	331	-3.9238	0.004322

ENSECAG00000022158	8	79696766	79722287	-13.8808	5.79E-25	1.36E-22	VPS4B (vacuolar protein sorting 4 homolog B)	3	-3.85375	0.003645
ENSECAG00000022647	9	6229159	6251198	12.97629	5.96E-22	7.12E-20	CHMP4C (charged multivesicular body protein 4C)	1	-3.56699	0.008631
ENSECAG00000021363	9	62858588	62998122	4.499673	2.56E-09	4.89E-08	DEPTOR (DEP domain containing MTOR-interacting protein)	12	-3.81331	0.005082
ENSECAG00000013971	10	56408907	56520676	10.3044	8.11E-14	2.47E-12	ATG5 (autophagy related 5)	110	-3.82991	0.005157
ENSECAG00000024499	10	58506453	58523225	1.941151	0.001093	0.008744	FOXO3 (forkhead box O3)	2	-3.62995	0.007238
ENSECAG00000019067	11	23985788	23999212	11.86848	1.25E-18	8.11E-17	OSBPL7 (oxysterol binding protein-like 7)	1	-3.55282	0.00899
ENSECAG00000015368	11	37580509	37583009	8.465334	1E-08	1.8E-07	PEX12 (peroxisomal biogenesis factor 12)	1	-3.577	0.008385
ENSECAG00000017425	11	51672958	51741179	-8.73255	2.9E-09	5.47E-08	PIK3R5 (phosphoinositide-3-kinase, regulatory subunit 5)	26	-3.75586	0.005367
ENSECAG00000014725	14	40760991	40922273	12.63797	9E-21	8.86E-19	PITX1 (paired-like homeodomain 1)	2	-3.57615	0.008405
ENSECAG00000007518	14	47635117	47670086	-2.68765	0.000318	0.003068	MARCH3 (membrane-associated ring finger (C3HC4) 3, E3 ubiquitin protein ligase)	124	-3.71976	0.006431
ENSECAG00000007942	14	47692226	47739561	-2.28296	0.000966	0.007839	LMNB1 (lamin B1)	116	-4.46256	0.000852
ENSECAG00000011406	14	68853182	68942841	8.651716	5.05E-09	9.27E-08	ST8SIA4 (ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4)	65	-3.80435	0.005286
ENSECAG00000000903	14	84754693	84910772	2.862874	0.000133	0.001434	RASGRF2 (Ras protein-specific guanine nucleotide-releasing factor 2)	23	-3.80045	0.005269
ENSECAG00000020847	14	86477973	86642232	2.519297	0.000189	0.001948	ARSB (arylsulfatase B)	30	-3.72394	0.005929
ENSECAG00000004289	15	10737295	10797292	9.128406	4.49E-11	9.99E-10	INPP4A (inositol polyphosphate-4-phosphatase, type I, 107kDa)	7	-3.64432	0.007787
ENSECAG00000009610	15	85914791	85956240	12.27764	1.26E-19	1.03E-17	RNF144A (ring finger protein 144A)	89	-3.89837	0.004025
ENSECAG00000000689	16	17404324	17450056	10.9353	8.94E-16	3.59E-14	PPP4R2 (protein phosphatase 4, regulatory subunit 2)	23	-3.89151	0.003803

ENSECAG00000016131	16	20621610	20787495	-5.07897	1.36E-10	2.88E-09	FRMD4B (FERM domain containing 4B)	3	-3.53354	0.009504
ENSECAG00000016372	17	4573937	4629149	10.45843	2.27E-14	7.56E-13	SPATA13 (spermatogenesis associated 13)	1	-3.60592	0.007709
ENSECAG00000015413	17	8971379	8991975	-3.26977	5.65E-06	8.1E-05	SLC7A1 (solute carrier family 7 member 1)	5	-3.61543	0.007614
ENSECAG00000000879	17	21719935	21817311	14.50011	2.9E-26	1.01E-23	CAB39L (calcium binding protein 39-like)	731	-4.17556	0.002171
ENSECAG00000002755	18	47662359	47663466	9.868204	8.61E-13	2.33E-11	B3GALT1 (UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 1)	4	-3.64494	0.007823
ENSECAG00000020788	19	9611359	9672717	10.36718	3.75E-14	1.21E-12	MECOM (MDS1 and EVI1 complex locus)	12	-4.02371	0.003176
ENSECAG00000001441	19	33563993	33586434	2.255299	0.000312	0.003015	TNK2 (tyrosine kinase, non-receptor, 2)	12	1.819854	0.007972
ENSECAG00000024988	19	36754148	36795593	-3.1029	1.98E-05	0.000262	PARP14 (poly (ADP-ribose) polymerase family, member 14)	188	-4.43196	0.001377
ENSECAG00000020093	21	31604725	31669820	3.353725	8.86E-05	0.000995	NPR3 (natriuretic peptide receptor C/guanylate cyclase C)	6	-3.60641	0.008081
ENSECAG00000023860	22	20308216	20347901	-11.9472	2.17E-18	1.35E-16	TGM3 (transglutaminase 3)	8	-3.84544	0.004061
ENSECAG00000013881	22	29084472	29249586	2.514713	0.000179	0.001853	DHX35 (DEAH (Asp-Glu-Ala-His) box polypeptide 35	81	-3.87529	0.004104
ENSECAG00000014676	22	36998574	37106417	-10.6894	7.17E-15	2.57E-13	PREX1 (phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 1)	4	1.750548	0.009563
ENSECAG00000008989	23	51439380	51584439	-9.40571	4.76E-12	1.16E-10	AUH (AU RNA binding protein/enoyl-CoA hydratase)	5	-3.74096	0.005355
ENSECAG00000003862	24	33675191	33720018	9.35181	3E-11	6.79E-10	C14orf102 (UPF0614 protein C14orf102-like)	3	-3.66206	0.006643
ENSECAG00000024297	25	18941561	19101286	-2.7057	0.000203	0.002075	COL27A1 (collagen, type XXVII, alpha 1)	5	-3.56944	0.008613
ENSECAG00000012050	26	32991276	33237308	10.65909	5.26E-15	1.91E-13	HLCS (holocarboxylase synthetase)	5	-3.82705	0.004542
ENSECAG00000015894	27	34212375	34265602	-2.93484	2.76E-05	0.000353	ANGPT2 (angiopoietin 2)	97	2.22192	0.002753
ENSECAG00000003554	29	29291814	29303208	14.40568	7.89E-26	2.42E-23	AKR1E2 (aldo-keto reductase family 1, member E2)	1	1.76432	0.009193

ENSECAG00000021968	X	18103355	18106014	-2.69849	7.55E-05	0.000861	SAT1 (spermidine/spermine N1-acetyltransferase 1)	13	1.790818	0.008567

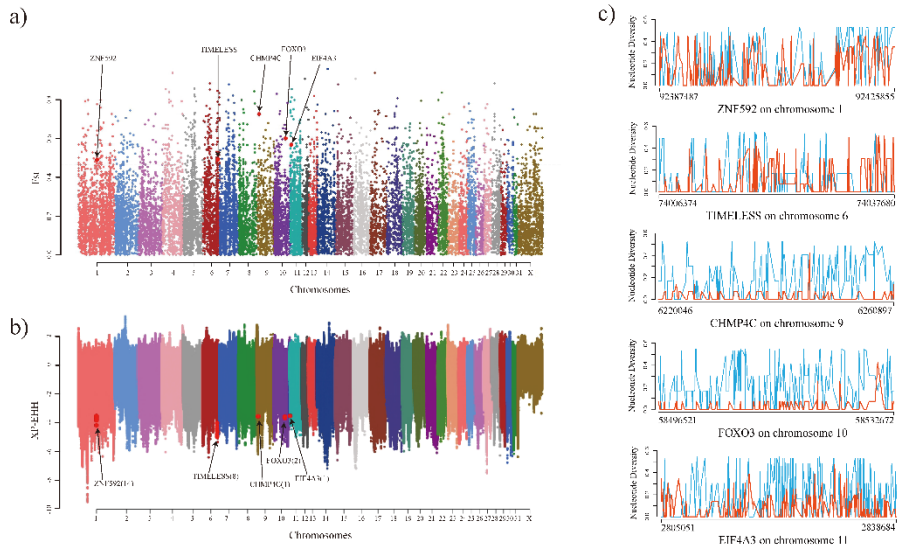


Figure 4.10. Signatures of correlation between DEGs from Thoroughbred RNA-seq and selected genes associated with nucleotide diversity, F_{ST} and XP-EHH from Thoroughbred and Jeju pony DNA sequence.

a) Manhattan plot of F_{ST} (Dotted line = cut-off value of the top 5% with empirical p-values of < 0.05. Red point = common genes between the DEGs, Thoroughbred selected genes associated with F_{ST} and XP-EHH). b) Manhattan plot of XP-EHH value (Red point = Common genes between the DEGs and Thoroughbred selected genes associated with F_{ST} and XP-EHH). c) Nucleotide diversity line plot of five common genes (sky blue color line = Jeju pony, orange color line = Thoroughbred).

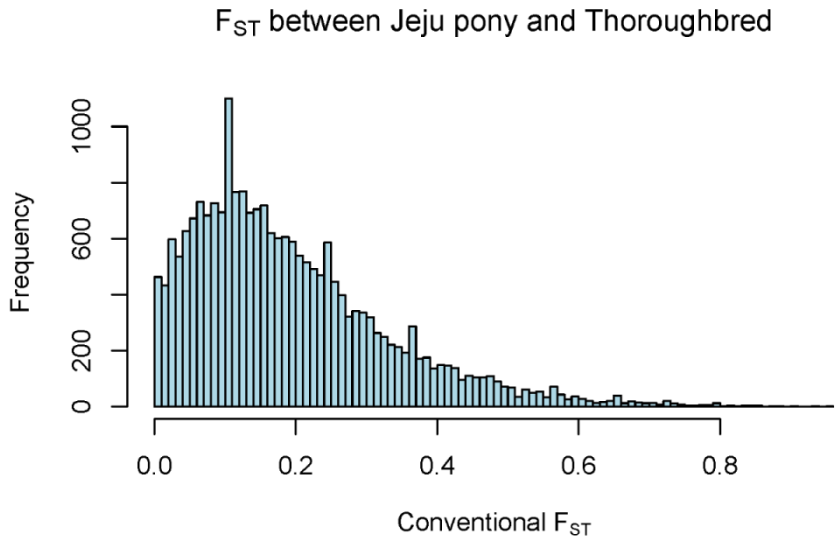


Figure 4.11. Histogram of conventional F_{ST} frequency between Thoroughbred and jeju pony.

x-axis is conventional F_{ST} value, y-axis is gene frequency.

4.5 Discussion

4.5.1 Differences in result between reference-based and de novo-based assembly and analysis

Up to now, many studies of RNA-seq analyses have used reference-based analysis (RBA) when a reference genome for the species is available (Garber et al. 2011, Peng et al. 2012, Kim et al. 2013). When a species does not have a reference genome, RBA using the reference genome of a closely related species or de novo-based analysis (DBA) is used. Several studies of RNA-seq analyses have used the align-then-assemble combined method (align-then-assemble and assemble-then-align) (Park et al. 2012). DBA used in align-then-assemble method assembles the unmapped sequence reads after RBA, which supplements the inherent weakness of RBA. However, I suggest that even if a species has a reference genome, DBA used in assemble-then-align is necessary for assembly of the total sequence reads including the unmapped sequence reads (Table 4.3 and Table 4.4). In this study, I found a significant difference in the number of transcript and differentially expressed genes (DEGs) identified by RBA and DBA. Also, a greater number of unique transcripts were identified by DBA than in RBA (Figure 4.1a and Figure 4.2a). This implies that the horse RNA-seq data in this study includes new transcripts and new transcript structures that are not included in the horse reference genome. Additionally, I found *de novo* unique differentially expressed genes (DUDEGs) which cannot be attained from RBA (Figure 4.1c

and Figure 4.2c). Credibility of DBA in RNA-seq has been proven in numerous methods and protocol papers (Grabherr et al. 2011, Martin et al. 2011, Henschel et al. 2012) and in this study. Both multidimensional scaling (MDS) plot (Figure 4.1b, Figure 4.2b and Figure 4.3) and expression patterns of common DEGs did not show differences between the results of RBA and DBA. However, the intensity of the expression was different because the newly assembled transcriptome reference represents the individual transcriptome made using de Bruijn graph assemblers during DBA (Martin et al. 2011) (Figure 4.1d and Figure 4.2d).

RNA-Seq can reveal sequence variations such as SNP in genes (Barbazuk et al. 2007) as is possible with whole genome sequencing (WGS). Transcribed SNPs in RNA-seq are needed for accurate measurement of allele-specific expression (Bray et al. 2003, Gregg et al. 2010) and detection of novel SNPs. (Xia et al. 2011). Hence, I compared the number and rate of SNPs identified from the two NGS methods and found differences between the type of references and NGS methods. In summary, I detected more SNPs in de novo assembly of RNA-seq than in the reference genome assembly of cDNA (Table 4.2).

4.5.2 Identification and Functional annotation of unique DEGs identified by de novo base assembly

I identified DUDEGs to ascertain the important function of DEGs, which cannot be attained from RBA. In the highest biological process gene ontology

(BP-GO) of DUDEG result, immune system process had the most significant P-value (P-value = 2.23E-13) in up-regulation of skeletal muscle (Figure 4.5). Response to stimulus had the second most significant P-value (P-value = 5.16E-08), which is related with immune response caused by exercise-induced stress (Kingston et al. 1996). Exercise-induced stress is closely related with the regulation of immune response (McGivney et al. 2010, Kim et al. 2013). Over-exercise in horses has shown an increase in the expression of alpha-1-antitrypsin protein, which plays an important role in protection of cells from inflammatory enzymes released from neutrophils (Stefansson et al. 2004). Exercise-induced reactive oxygen species was also related with the regulation of immune responses, and caused the inflammatory responses from muscle damage (Niess et al. 1999, Dousset et al. 2007). In the KEGG pathways result (Table 4.6), the JAK-STAT signaling pathway and MAPK signaling pathway were also up-regulated in the skeletal muscle. The JAK-STAT signaling is a key pathway in myoblast proliferation (Sun et al. 2007) and plays a major role in inflammatory and immune responses (O'Shea et al. 2004). The MAPK signaling pathway is implicated in inflammation and carbohydrate metabolism (Chau Long et al. 2004). Death, related with apoptosis of skeletal muscle caused by over-exercise also had a significant P-value (P-value = 1.75E-06). This was supported by the KEGG pathways results, which showed that the p53 signaling pathway and regulation of actin cytoskeleton were up-regulated. P53 protein has an important role in apoptosis of skeletal muscle and actin cytoskeleton, and is also a key pathway in regulation of apoptosis pathways (Gourlay et al. 2005, Saleem et al. 2009). However, I did not find immune responses and apoptosis related with BP-GO and KEGG pathway in blood.

4.5.3 Integration of conceptually new DEGs: DUDEGs and selected gene associated with nucleotide diversity, F_{ST} and Cross Population Extended Haplotype Homozygosity (XP-EHH)

In order to investigate the evolutionary history of domestication in relation to different experimental conditions, I approached the identification of DEGs with a new concept. The conceptually new DEGs were attained by screening for genes in common between DUDEGs from DBA in RNA-seq and selected genes identified by evolutionary statistics, such as nucleotide diversity, F_{ST} and XP-EHH from RBA in WGS.

This comparison highlighted three genes (*EIF4A3, ZNF592* and *TIMELESS*) in blood and two genes (*CHMP4C* and *FOXO3*) in skeletal muscle as being in common between DUDEGs, F_{ST} and XP-EHH. These five genes are not only DUDEGs in six Thoroughbreds, before and after exercise, but also selected genes (F_{ST} (empirical p-value < 0.01) and XP-EHH (value < 0 and p-value < 0.01)). A pairwise test, XP-EHH, of the Thoroughbred and Jeju domestic pony populations was used to identify selective sweep regions between the two populations. As I am interested in locating selective sweep regions representing adaptation in Thoroughbreds, a cutoff of XP-EHH value < 0 was used. If the XP-EHH value > 0 is used, then the identified selective sweep region would correspond to the adaptations in Jeju domestic pony.

The five genes were conceptually new DEGs, and were related to the evolution of exercise response during the domestication process of

Thoroughbred. Among them, three genes, *CHMP4C*, *EIF4A3* and *FOXO3*, showed relatively low levels of nucleotide diversity compared to that of the Jeju pony (Table 4.7 and Figure 4.10). This suggests that these three genes have been more strongly selected for in Thoroughbred than in Jeju pony. *EIF4A3* was mostly expressed in megakaryocytes, platelets and red blood cell. *EIF4A3*, an mRNA-localization protein in mammals, controls the synaptic strength, neuronal protein expression, and in megakaryocytes and platelets act as mRNA sorting machinery (Giorgi et al. 2007, D'Alessandro et al. 2009, Cecchetti et al. 2011). In a previous study, it was shown that over-exercise activates and increases platelets (Kestin et al. 1993). Although, it has an important role in blood post-exercise in Thoroughbreds, *EIF4A3* expression was up-regulated in my results. *CHMP4C* is a p53-regulated gene and plays an important role in exosome production (Yu et al. 2009). The importance of p53 in apoptosis of skeletal muscle was implicated in a previous study, in which p53-null animals showed greater fatigability and less locomotory endurance than wild-type animals (Saleem et al. 2009). This suggests that p53 is closely related with exercise-induced stress in skeletal muscle. *CHMP4C* expression was down-regulated in my results implying the activation of p53 regulation in Thoroughbred skeletal muscle. *FOXO3*, also known as Forkhead box O3, has a role in triggering apoptosis by down-regulating the *FOXO3* gene. In addition, *FOXO3* causes a loss of muscle mass, and is closely related to *PGC1α, ATG4b, ATG12, Beclin1*, *Gabarapl1*, and *LC3b*. *PGC1α*, the transcription of atrophy-specific genes, inhibits the activity of the transcription factor *FOXO3*, with protects skeletal muscle from atrophy (Sandri et al. 2006). In human muscle after ultra-endurance exercise, the

expression of several autophagy genes, *ATG4b, ATG12, Beclin1, Gabarapl1* and *LC3b*, were increased(Zhao et al. 2007, Jamart et al. 2012). For this reason, *FOXO3* was also closely related to exercise in Thoroughbred skeletal muscle.

Based on these results, *EIF4A3, CHMP4C* and *FOXO3* are conceptually new DEGs involved in exercise response that have been selected for during the domestication history of the Thoroughbred that cannot be acquired by RBA.

This chapter will be published in elsewhere
as a partial fulfillment of Woncheoul Park's Ph.D program.

Chapter 5. Differentially expressed isoform, splicing and an alternative splicing event frequency in Thoroughbred race horses before and after exercise

5.1 Abstract

In this study, I aim to identity that differentially expressed isoforms (DEIs), differential splicing and alternative splicing event by using the published Thoroughbred racing horse RNA-seq data between before and after exercise, because previous studies didn't researched that carefully and without researches about alternative splicing event in Thoroughbred racing horses. I used g/--GTF-guide option in Cufflinks program, because I wants to find the all reference transcripts as well as any novel genes, isoform and splicing.

As a result, In DEIs, the number of DEI in blood and skeletal muscle were 67 and 1,133 respectively. Among them, novel DEIs were 37 in blood, 378 in skeletal muscle. In addition, I identified 7 (6 up-regulated and 1 down-regulated) DEIs in blood and 56 (45 up-regulated and 11 down-regulated) DEIs in skeletal muscle. Among them, in blood, 3 isoforms such as HSPA8 (heat shock 70 kDa protein 8 gene), *RhoB* (Rho-related GTP-binding protein) and *SOCS3* (suppressor of cytokine signaling 3 mRNA) (up-regulated) in blood and 5 isoforms such as *AMPD*2 (AMP Deaminase Isoform L), *ICAM1* (intercellular adhesion molecule 1), *MMP-1* (Matrix metalloproteinase-1), *MXD1* (MAX Dimerization Protein 1) and *TET2* in skeletal muscle were revealed that related to exercise-induces. Moreover, I identified 4 (4 up-regulated) significant differential splicing such as *BLZF1, ITGB6, KDM5C* and *ZNF207* gene in skeletal muscle. Most of these genes were included a litter-related exercise-induce stress with alternative splicing. Conclusively, I classified and identified the alternative splicing events in blood and skeletal

muscle in six Thoroughbreds racing horses before and after exercise. As a result, I identified that exon skipping/inclusion (ESI) type is the most common of alternative splicing event, this is the identical result such as human and yeast but the different result as pig with alternative 3' splicing (A3)

5.2 Introduction

Thoroughbred racing horses are a higher value-added animal. Because over 40 countries have been performing Thoroughbred horse racing and including over half a million horses worldwide since Tudor times (Gaffney et al. 1988). In addition, the British and Irish Thoroughbred horse racing and breeding industry were a multimillion pound concern in 1982 (Jeffcott et al. 1982) but these industry is thought to be much higher. Previously, most researcher studied the analysis of Thoroughbred phenotype, nutrition, injury, condition and racing performance by using statistical method and simply experiment method (Lindner et al. 1993, Murray et al. 1996, Johnson et al. 2001, White et al. 2001, Takahashi et al. 2004). For this reason, at present, next generation sequencing (NGS) data in Thoroughbred racing horses are valuable asset to researchers with studying to Thoroughbred racing horses so they used the public NGS data, for example, whole genome and transcriptome analysis.

In this study, used Thoroughbred racing horses RNA-seq data were public data that used in 3 papers (Jung et al. 2012, Kim et al. 2013, Park et al. 2014), but these papers did not go into detail about isoform and splicing such as differentially expressed isoform, differential splicing and alternative splicing. Pre-mRNA splicing in which introns are removed and exons are joined is play an important step in eukaryotic gene expression. In the spliceosome, the series of reactions of pre-mRNA splicing appear. Conserved splicing signals in the pre-mRNA defined the site (including the 5’ splice site and the 3’ splice site) for cleavage and ligation in splicing reactions (Burge et

al. 1999). Alternative splicing is also play an important role in regulation of gene activity in eukaryotic species. The many alternative splicing event have been reported, most common alternative splicing events are exon inclusion or skipping, intron removal or retention and alternative 5' and 3' splice event (Wu et al. 2004). many alternative splicing events are very uncommon, appear in a specific tissue at a specific time in physiological conditions and/or development(Graveley 2001). These alternative splicing frequency was reported in some eukaryotic species such as human, mouse and pig (Mironov et al. 1999, Pan et al. 2005, Chen et al. 2011), however unreported in Thoroughbred racing horses.

In this paper, I describe an alternative splicing events frequency, differentially expressed isoforms and differential splicing in blood and skeletal muscle RNA-seq data from Thoroughbred racing horses before and after exercise.

5.3 Materials and methods

5.3.1 Animals

Thoroughbred racing horse RNA-seq data were collected from public release and were available free of charge from the NCBI Gene Expression Omnibus (GSE37870). In samples, Twenty-four sets of transcriptome data were collected from muscle and blood samples of six Thoroughbred racing horses obtained before and after exercise.

5.3.2 Mapping of RNA-seq reads

I trimmed the adapt sequence, the specific sequence of the other ILLUMINA and below 80bp reads by Trimmomatic program (version 0.32, option: ILLUMINACLIP:TruSeq3-PE-2.fa :2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) (Bolger et al. 2014) before alignment. After then, I aligned the transcript reads to the ENSEMBL horse reference genome (*EquCab2*) (ftp://ftp.ensembl.org/pub/release-76/fasta/equus_caballus/dna/) using TopHat ver.2.0.12 program (Trapnell et al. 2009) that is a fast splice junction mapper for RNA-seq. TopHat contain the Bowtie2 ver.2.2.3 program (Langmead et al. 2012) to achieve the alignment. When I executed the program, I used the default option.

5.3.3 Transcript assembly and differential expression testing

I processed the aligned reads (TopHat output files) using Cufflinks ver.2.2.1 program (Trapnell et al. 2013) following the default option basically, except for the following option: -g/--GTF-guide that find novel genes and isoform by Reference Annotation Based Transcript (RABT) assembly. Cufflinks uses the normalized RNA-seq fragment to estimate the abundances of transcripts. Fragments Per Kilobase of exon per Million fragments mapped (FPKM) of each sample were count to estimate the expression levels of the transcripts. Cufflinks contain the Cuffdiff to estimate the differential expression of transcripts across condition points and finds significant changes in transcripts expression, splicing and promoter use. So I used Cuffdiff to achieve two pairwise comparisons of transcripts expression and splicing between before and after exercise in horse muscle and blood RNA-seq. significant differentially expressed genes (DEGs) and DEIs were identified by FDR < 0.05, significant differential splicing was identified by FDR < 0.1.

5.3.4 Visualization of mapped reads and Alternative splicing event

Mapping results were visualized using the University of California, Santa Cruz (UCSC) genome browser (Zweig et al. 2008) and a local copy of the Integrative Genomic Viewer (IGV) tool (http://www.broadinstitute.org/igv/). Alternative splicing event was visualized using spliceR tool (Vitting-Seerup et al. 2014)

5.3.5 Functional analysis of transcript lists using DAVID

Gene Ontology (GO) categories are commonly used in this technique and there are many tools available for performing GO and KEGG pathway analysis. I used DAVID web tool(Da Wei Huang et al. 2008) to convert the horse Ensembl transcripts IDs to official gene symbols. This was carried out by cross-matching the horse Ensembl transcript IDs to the human official gene symbols. The representation of functional groups in blood and skeletal muscle relative to the whole genome was investigated using the Expression Analysis Systematic Explorer (EASE) tool within DAVID (Hosack et al. 2003). The EASE tool is a modified Fisher's exact test used to measure enrichment of gene ontology terms. To identify enriched GO terms and KEGG pathway, functionally clustered gene symbols were filtered by an EASE value of < 0.1, except for the filtering EASE value of < 0.01: muscle up-regulated GO Biological Process (BP) terms.

5.4 Results

5.4.1 Mapping of the horse transcriptome

In each Thoroughbred racing horse sample transcriptome, about 0.5 ~ 1.8 % of reads were filter out with the result that using Trimmomatic program. And I aligned all short reads onto the whole horse reference genome (*EquCab2*). About 83.2 ~ 95.5% of reads were overall mapped and 78.1 ~ 91.1% of reads concordant pair mapped to the horse reference genome. In addition, multiple alignment rate was 3.3 ~ 8.3% (Table 5.1).

Table 5.1. Summary of RNA-seq reads and mapping rate of before and after exercise from six Thoroughbred blood and muscle

	BF						BS					
	1B	1P	2B	2P	3B	3P	1B	1P	2B	2P	3B	3P
Reads Pairs	26577778	26111110	26111110	26111110	26111110	26577778	26111110	25622222	26011111	25007819	25833299	26133334
Dropped-Trimmomatic	21966 (0.08%)	19764 (0.08%)	15692 (0.06%)	15689 (0.06%)	19465 (0.07%)	20176 (0.08%)	19156 (0.07%)	16046 (0.06%)	15549 (0.06%)	11701 (0.05%)	15510 (0.06%)	12275 (0.05%)
Overall Alignment rate	92.70%	92.40%	93.20%	93.30%	92.40%	92.60%	93.00%	92.60%	92.60%	83.20%	92.40%	91.50%
concordant pair alignment rate	88.00%	87.50%	88.50%	88.50%	87.70%	88.10%	87.90%	87.60%	87.50%	78.10%	87.10%	85.80%
Multiple alignment rate	3.40%	3.30%	3.60%	3.60%	3.60%	3.30%	3.80%	3.60%	3.80%	3.70%	3.90%	4.00%

	MF						MS					
	1B	1P	2B	2P	3B	3P	1B	1P	2B	2P	3B	3P
Reads Pairs	28953336	28218561	28606040	28606040	28953336	28595004	28953336	28953336	28953336	28069445	28069445	28069445
Dropped-Trimmomatic	19903 (0.07%)	19575 (0.07%)	49883 (0.17%)	49883 (0.17%)	47956 (0.17%)	42091 (0.15%)	41058 (0.14%)	37262 (0.13%)	46561 (0.16%)	50025 (0.18%)	41711 (0.15%)	45340 (0.16%)
Overall Alignment rate	93.50%	93.90%	93.90%	94.80%	94.60%	95.50%	92.90%	93.40%	93.50%	93.50%	93.90%	94.20%
concordant pair alignment rate	86.20%	86.60%	87.40%	88.40%	90.00%	91.00%	85.40%	86.00%	85.60%	85.50%	85.80%	85.50%
Multiple alignment rate	7.10%	7.00%	7.10%	7.20%	6.70%	6.50%	7.90%	7.50%	7.30%	7.70%	8.30%	7.10%

5.4.2 Differential expression analysis and Alternative splicing identification

As a result of Cuffdiff program running, I found the DEGs and the DEIs in blood and skeletal muscle RNA-seq data from six Thoroughbreds before and after exercise. In DEGs, the number of DEG in blood and skeletal muscle were 124 and 2,589 respectively. In DEIs, the number of DEI in blood and skeletal muscle were 67 and 1,133 respectively. Among them, novel DEGs and DEIs were 57 and 37 in blood, 684 and 378 in skeletal muscle respectively (Supplementary file 5.1). In addition, I identified 7 (6 up-regulated and 1 down-regulated) DEIs in blood and 56 (45 up-regulated and 11 down-regulated) DEIs in skeletal muscle, except for non-gene symbol and log2 (fold change) value as 'inf' in the DEGs and the DEIs. Moreover, I identified 4 (4 up-regulated) significant differential splicing such as *BLZF1, ITGB6, KDM5C* and *ZNF207* gene in skeletal muscle (Table 5.2). I visualized these splicing by using UCSC genome browser and IGV (Figure 5.2 and Figure 5.3).

I classified and identified the alternative splicing events by using spliceR tool. In addition, the number of all alternative splicing events and average number of alternative splicing events per transcript in the different alternative splicing events types (single and multiple exon skipping/inclusion (ESI, MESI), intron skipping/inclusion (ISI), alternative donor and acceptor sites (A5, A3), alternative first or last exon usage (ATSS, ATTS) and mutually exclusive exon events (MEE)) shown in Figure 5.1. In blood, the number of all alternative splicing events in before and after exercise were 107,067 and

107,234 respectively, the number of intersect alternative splicing events between before and after exercise was 107,005. In addition, the average number of specific alternative splicing events per transcript in before and after exercise were 0.4 and 0.99 respectively, the average number of intersect alternative splicing events between before and after exercise was 2.48. In muscle, the number of all alternative splicing events in before and after exercise were 67385 and 67382 respectively, the number of intersect alternative splicing events between before and after exercise was 67347. In addition, the average number of specific alternative splicing events per transcript in before and after exercise were 0.32 and 0.26 respectively, the average number of intersect alternative splicing events between before and after exercise was 2.12 (Figure 5.1).

Table 5.2. List of DEIs (FDR<0.01) in blood and skeletal muscle, and DS (FDR <0.1) in skeletal muscle in six Thoroughbred horses before and after exercise RNA-seq data by using Cuffdiff within Cufflinks

Differentially expressed isoforms in Muscle													
test_id	**gene_id**	**gene**	**locus**	**sample1**	**sample2**	**status**	**value_1**	**value_2**	**log2(fold_change)**	**test_stat**	**p_value**	**q_value**	**significant**
TCONS_00030235	XLOC_011456	41701	14:47634665-47677366	q1	q2	OK	1.45013	5.30907	1.87227	1.58457	0.00085	0.030591	yes
TCONS_00006180	XLOC_001881	AEN	1:94539488-94568306	q1	q2	OK	1.21839	5.65439	2.21439	2.74217	0.00005	0.002828	yes
TCONS_00112787	XLOC_045944	ALPK2	8:75684861-75805550	q1	q2	OK	0.931762	4.7156	2.33941	1.69379	0.00065	0.02493	yes
TCONS_00097495	XLOC_039282	AMPD2	5:58505198-58577882	q1	q2	OK	1.0182	8.29095	3.02551	2.54495	0.00005	0.002828	yes
TCONS_00038876	XLOC_014953	CMTM6	16:51678633-51706985	q1	q2	OK	2.35012	6.25418	1.41209	2.34704	0.00005	0.002828	yes
TCONS_00087119	XLOC_035163	CSRP1	30:28518065-28541752	q1	q2	OK	15.9524	33.6361	1.07623	1.84714	0.00005	0.002828	yes
TCONS_00103842	XLOC_041897	CTDSP2	6:75350109-75359905	q1	q2	OK	33.3603	19.5104	-0.77389	-1.40462	0.0005	0.020176	yes
TCONS_00064456	XLOC_025530	eca-mir-1-2	22:48427158-48456583	q1	q2	OK	5.29292	2.4708	-1.09908	-1.84278	0.00005	0.002828	yes
TCONS_00058180	XLOC_022870	eca-mir-206-2	20:49820594-49826407	q1	q2	OK	1.64425	0.51708	-1.66897	-1.69659	0.0005	0.020176	yes
TCONS_00017339	XLOC_006823	eca-mir-21	11:33729763-33866347	q1	q2	OK	4.57788	80.308	4.13279	4.52247	0.00005	0.002828	yes
TCONS_00123900	XLOC_051067	eca-mir-221	X:37089636-37091979	q1	q2	OK	0.181817	15.2214	6.38747	4.90477	0.00005	0.002828	yes
TCONS_00029982	XLOC_011382	ECSCR	14:37210755-37235258	q1	q2	OK	14.7476	28.7828	0.964724	1.59947	0.00015	0.007496	yes
TCONS_00054339	XLOC_021032	EFHD2	2:37427563-37443966	q1	q2	OK	0.603103	2.66075	2.14136	2.29845	0.00005	0.002828	yes
TCONS_00018063	XLOC_007009	EIF4A1	11:50544621-50561301	q1	q2	OK	20.2136	109.664	2.4397	2.66194	0.00005	0.002828	yes

TCONS_00113104	XLOC_046032	ENSECAG00000002100	8:1480709-1495491	q1	q2	OK	12.5206	6.01278	-1.0582	-1.39178	0.00145	0.046294	yes
TCONS_00061224	XLOC_024288	ENSECAG00000006264	21:12165105-12174134	q1	q2	OK	0.320454	0.711618	1.15098	1.42443	0.0015	0.047338	yes
TCONS_00101713	XLOC_041288	ENSECAG00000006711	6:7827180-7882939	q1	q2	OK	0.351557	3.49828	3.31482	2.17183	0.0001	0.005194	yes
TCONS_00094631	XLOC_038456	ENSECAG00000008084	5:47104961-47115530	q1	q2	OK	1.14213	3.53118	1.62843	2.39984	0.00005	0.002828	yes
TCONS_00005076	XLOC_001557	ENSECAG00000010185	1:38949379-39015338	q1	q2	OK	5.137	15.5988	1.60244	2.05768	0.00005	0.002828	yes
TCONS_00108086	XLOC_043806	ENSECAG00000012876	7:14573215-14608847	q1	q2	OK	1.04248	8.22149	2.97938	2.54369	0.00005	0.002828	yes
TCONS_00064201	XLOC_025476	ENSECAG00000019765	22:38285616-38308628	q1	q2	OK	3.58825	16.7078	2.21917	3.12718	0.00005	0.002828	yes
TCONS_00038101	XLOC_014773	ENSECAG00000021806	16:34935935-34957026	q1	q2	OK	13.5581	129.61	3.25695	1.87068	0.00005	0.002828	yes
TCONS_00087011	XLOC_035144	ENSECAG00000025915	30:26395642-26399027	q1	q2	OK	1.77014	0.470127	-1.91274	-1.70225	0.00055	0.021726	yes
TCONS_00061609	XLOC_024429	FAM134B	21:43901006-43996209	q1	q2	OK	10.9857	31.781	1.53253	1.67736	0.001	0.034761	yes
TCONS_00069109	XLOC_027674	FCF1	24:20169393-20247623	q1	q2	OK	3.16461	8.59568	1.44158	1.6539	0.00045	0.018514	yes
TCONS_00001819	XLOC_000523	FURIN	1:93025311-93036102	q1	q2	OK	1.64858	4.43953	1.42918	1.32773	0.00105	0.035973	yes
TCONS_00089075	XLOC_036126	GNG11	4:37525081-37532289	q1	q2	OK	8.64816	3.36609	-1.36132	-1.88668	0.00005	0.002828	yes
TCONS_00108447	XLOC_043910	HSPA8	7:30252002-30256636	q1	q2	OK	318.492	670.663	1.07433	1.56732	0.0003	0.013311	yes
TCONS_00109033	XLOC_044118	ICAM1	7:49683343-49692829	q1	q2	OK	0.47276	20.6926	5.45186	4.65508	0.00005	0.002828	yes
TCONS_00109035	XLOC_044118	ICAM1	7:49683343-49692829	q1	q2	OK	0.837544	21.1083	4.6555	3.70072	0.00005	0.002828	yes
TCONS_00106323	XLOC_043257	ICAM3	7:49621646-49668924	q1	q2	OK	3.44032	23.2409	2.75605	1.78323	0.00005	0.002828	yes
TCONS_00038100	XLOC_014773	ITIH4	16:34935935-34957026	q1	q2	OK	83.6305	452.636	2.43625	2.10496	0.00005	0.002828	yes
TCONS_00069100	XLOC_027671	LIN52	24:19716239-19828272	q1	q2	OK	6.62465	3.13261	-1.08048	-1.54334	0.00025	0.011523	yes

TCONS_00083795	XLOC_033550	MAF	3:27647698-27664054	q1	q2	OK	29.6611	15.3714	-0.948323	-1.52491	0.00065	0.02493	yes
TCONS_00107274	XLOC_043585	MICAL2	7:81016727-81154587	q1	q2	OK	1.37845	4.41088	1.67802	1.30822	0.00085	0.030591	yes
TCONS_00108057	XLOC_043796	MMP-1	7:12601841-12693969	q1	q2	OK	0.331967	18.7165	5.81713	3.92839	0.001	0.034761	yes
TCONS_00033888	XLOC_013118	MOB1A	15:28947416-28965967	q1	q2	OK	4.75405	12.1624	1.35519	2.12527	0.00005	0.002828	yes
TCONS_00039187	XLOC_015046	MRAS	16:73025185-73091969	q1	q2	OK	1.63457	5.12269	1.64799	2.22152	0.00005	0.002828	yes
TCONS_00081218	XLOC_032784	MT1F	3:8998214-9041556	q1	q2	OK	58.568	252.55	2.10838	2.38524	0.00005	0.002828	yes
TCONS_00035759	XLOC_013626	MXD1	15:32548522-32591783	q1	q2	OK	1.79335	10.9702	2.61286	1.70719	0.00045	0.018514	yes
TCONS_00083180	XLOC_033370	MXD4	3:117944455-117960975	q1	q2	OK	5.19	1.8013	-1.52669	-2.15683	0.00005	0.002828	yes
TCONS_00078139	XLOC_031316	MYF6	28:7861299-8072823	q1	q2	OK	7.91611	22.4121	1.50141	1.49775	0.0008	0.029235	yes
TCONS_00007044	XLOC_002174	NMES1	1:144015842-144022106	q1	q2	OK	0.584043	4.89822	3.06811	2.70504	0.00005	0.002828	yes
TCONS_00082620	XLOC_033208	PPAT	3:76335698-76409191	q1	q2	OK	4.29693	16.2229	1.91665	2.24363	0.00005	0.002828	yes
TCONS_00004154	XLOC_001298	PTGDR	1:185704379-185719951	q1	q2	OK	0.894162	2.51191	1.49018	1.69872	0.0002	0.009501	yes
TCONS_00112228	XLOC_045752	PXMP2	8:29646191-29657148	q1	q2	OK	10.3624	5.19183	-0.997047	-1.36335	0.00125	0.041508	yes
TCONS_00075892	XLOC_030488	RAB11FIP1	27:7747900-7789957	q1	q2	OK	0.171873	2.31396	3.75094	2.29659	0.00085	0.030591	yes
TCONS_00031392	XLOC_011825	SH3RF2	14:31634766-31792154	q1	q2	OK	0.124897	3.76787	4.91494	0.95032	0.0014	0.044979	yes
TCONS_00095772	XLOC_038792	SLC19A2	5:5823884-5919055	q1	q2	OK	4.83729	31.4314	2.69994	3.24172	0.00005	0.002828	yes
TCONS_00021183	XLOC_007754	SNORD65	11:57776309-57843628	q1	q2	OK	0.702944	3.86945	2.46065	1.72422	0.0003	0.013311	yes
TCONS_00090892	XLOC_036738	STEAP4	4:32515585-32545753	q1	q2	OK	0.881908	2.6101	1.5654	1.88817	0.00005	0.002828	yes
TCONS_00055652	XLOC_021435	TET2	2:119538122-119656413	q1	q2	OK	1.45554	7.42678	2.35118	2.66109	0.00005	0.002828	yes

TCONS_00061733	XLOC_024475	TPPP	21:57263130-57280281	q1	q2	OK	3.85507	2.10305	-0.874273	-1.46734	0.0006	0.023364	yes
TCONS_00007076	XLOC_002182	TRIM69	1:144406584-144487707	q1	q2	OK	0.74272	2.25736	1.60375	1.61872	0.0002	0.009501	yes
TCONS_00103313	XLOC_041725	TUBA1A	6:66914515-66961157	q1	q2	OK	13.1344	30.3136	1.20662	1.70784	0.00005	0.002828	yes
TCONS_00095706	XLOC_038771	YOD1	5:3236647-3269689	q1	q2	OK	0.826018	24.1738	4.87113	3.47542	0.0003	0.013311	yes
Differentially expressed isoforms in Blood													
TCONS_00102648	XLOC_041534	ENSECAG00000017267	6:37436304-37449129	q1	q2	OK	32.1515	56.783	0.820573	1.63202	0.00005	0.038842	yes
TCONS_00112537	XLOC_045848	HRH4	8:46199342-46225727	q1	q2	OK	12.2987	6.28861	-0.967692	-1.66659	0.00005	0.038842	yes
TCONS_00108447	XLOC_043910	HSPA8	7:30252002-30256636	q1	q2	OK	316.384	630.43	0.99466	2.1401	0.00005	0.038842	yes
TCONS_00103722	XLOC_041874	PTGES3	6:74201856-74223948	q1	q2	OK	6.50598	19.6583	1.5953	1.8698	0.00005	0.038842	yes
TCONS_00036581	XLOC_013852	RHOB	15:74749329-74764961	q1	q2	OK	17.9356	38.0128	1.08366	2.82718	0.00005	0.038842	yes
TCONS_00080548	XLOC_032330	SNORD112	29:27744816-27755397	q1	q2	OK	0.418002	0.979769	1.22893	2.14008	0.00005	0.038842	yes
TCONS_00015943	XLOC_006471	SOCS3	11:4215262-4217457	q1	q2	OK	14.831	23.4865	0.663217	1.75377	0.00005	0.038842	yes
Differential splicing in Muscle													
TSS12403	XLOC_007517	ZNF207	11:40110966-40183336	q1	q2	OK	0	0	0.671304	0	5.00E-05	0.069388	yes
TSS28812	XLOC_017961	ITGB6	18:40788335-40917088	q1	q2	OK	0	0	0.560692	0	5.00E-05	0.069388	yes
TSS59280	XLOC_038060	BLZF1	5:5798139-5822733	q1	q2	OK	0	0	0.284047	0	5.00E-05	0.069388	yes
TSS78142	XLOC_050348	KDM5C	X:45020745-45053199	q1	q2	OK	0	0	0.390994	0	5.00E-05	0.069388	yes

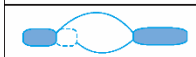
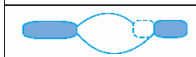
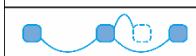
Alternative splicing Event type		Blood						Muscle					
		Before After		Before After		Before After		Before After		Before After		Before After	
		No. of events	Ave No. of events per ranscript	No. of events	Ave No. of events per ranscript	No. of events	Ave No. of events per ranscript	No. of events	Ave No. of events per ranscript	No. of events	Ave No. of events per ranscript	No. of events	Ave No. of events per ranscript
	Exon skipping/inclusion (ESI)	11	0.07	27823	0.64	50	0.22	7	0.06	18363	0.58	2	0.01
	Multiple exon skipping/inclusion (MESI)	3	0.2	6809	0.16	10	0.04	0	0.2	4120	0.13	0	0.04
	Intron skipping/inclusion (ISI)	0	0	8700	0.2	3	0.01	5	0.04	4225	0.13	1	0.01
	Alternative 5' splice sites (A5)	1	0.01	10787	0.25	21	0.09	3	0.03	6046	0.19	4	0.03
	Alternative 3' splice sites (A3)	4	0.03	17260	0.4	48	0.21	2	0.02	10061	0.32	5	0.04
	Alternative transcription start site (ATSS)	20	0.13	24868	0.58	45	0.19	10	0.08	16970	0.53	12	0.09
	Alternative transcription termination site (ATTS)	23	0.15	10620	0.25	42	0.18	11	0.09	7432	0.23	11	0.08
	Mutually exclusive exon (MEE)	0	0	138	0	10	0.04	0	0	130	0	10	0
	All events	62	0.19	107005	1.86	229	0.16	38	0.32	67347	2.12	35	0.26

Figure 5.1 Number of individual alternative splicing events and average number of alternative splicing events per transcript identified

A schematic representation of the each alternative splicing type, the number of classified events and average number of classified event per transcript in blood and skeletal muscle RNA-seq data from Thoroughbred horse before and after exerci

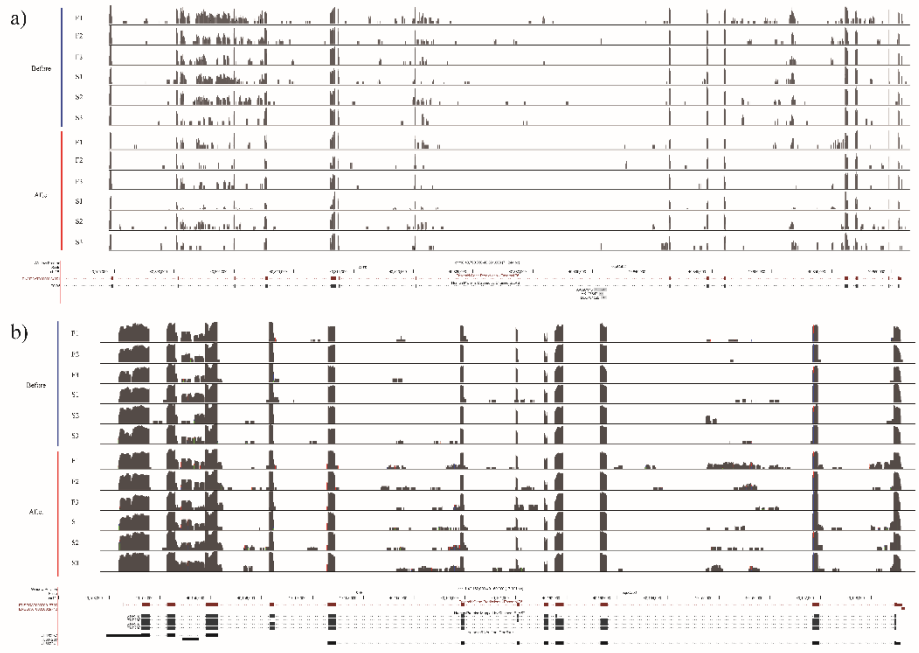


Figure 5.2. RNA-seq read mapping to the horse reference for differentially splicing such as ITGB6 and ZNF207.

RNA-Seq read mapping to the ENSEMBL horse reference genome (EquCab2) of the genes a) ITGB6 and b) ZNF207 for all 12 samples in skeletal muscle RNA-seq data from Thoroughbred horse before and after exercise. The before exercise condition are shown in y axis blue bar and after exercise condition samples in y axis red bar A schematic representation of the Ensembl transcripts for differentially splicing is shown in brown, human gene symbol is shown in dark grey and horse mRNA from GeneBank is shown black at the bottom of the figure. (*'F1', 'F2', 'F3', S1', S2' and 'S3' are the name of the horse samples).

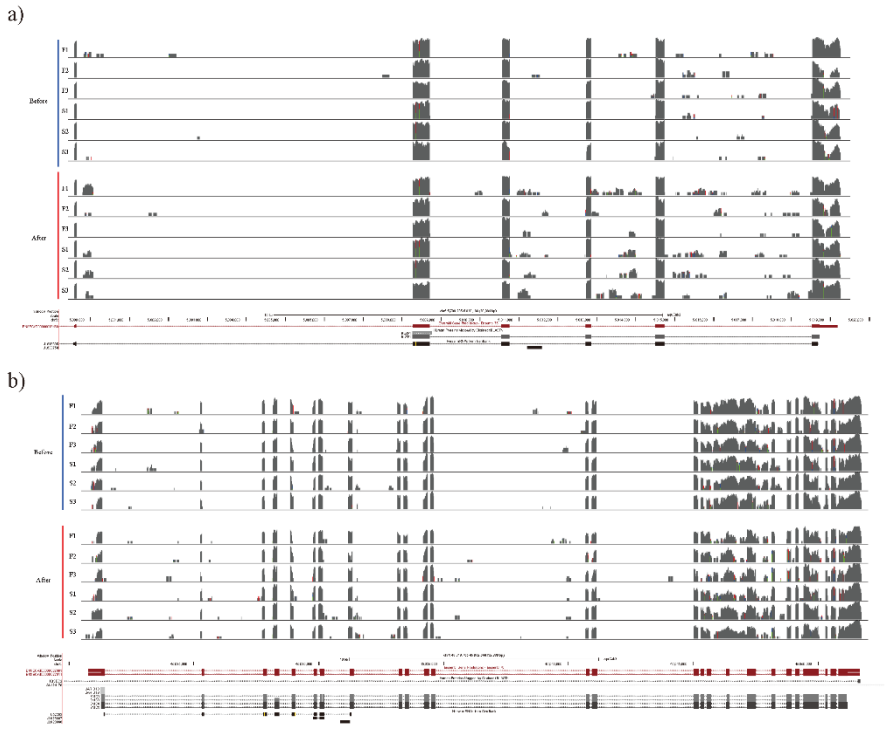


Figure 5.3. RNA-seq read mapping to the horse reference for differentially splicing such as BLZF1 and KDM5C.

RNA-Seq read mapping to the ENSEMBL horse reference genome (EquCab2) of the genes BLZF1 and KDM5C for all 12 samples in skeletal muscle RNA-seq data from Thoroughbred horse before and after exercise. The before exercise condition are shown in y axis blue bar and after exercise condition samples in y axis red bar A schematic representation of the Ensembl transcripts for differentially splicing is shown in brown, human gene symbol is shown in black and horse mRNA from GeneBank is shown black at the bottom of the figure. (*'F1', 'F2', 'F3', S1', S2' and 'S3' are the name of the horse samples).

5.4.3 Functional annotation of DEIs

I summarized the highest gene ontology biological process (GOBP) term of DEIs in blood and skeletal muscle taken before and after exercise from six Thoroughbred RNA-seq data (Figure 5.4). In blood, The most significantly enriched terms in up-regulated isoforms were ‘Biological regulation’, ‘Death’, ‘Developmental process’, ‘Multicellular organismal process’ and ‘Response to stimulus’ but down-regulated isoforms were no GOBP terms. In skeletal muscle, The most significantly enriched terms in up-regulated isoforms were ‘Biological regulation’, ‘Cellular component organization’, ‘Cellular process’, ‘Death’, ‘Developmental process’, ‘Growth’, ‘Immune system process’, ‘Localization’, ‘Locomotion’, ‘Multicellular organismal process’, ‘Multi-organism process’, ‘Response to stimulus’, ‘Rhythmic process’ and ‘Viral reproduction’, moreover ‘Developmental process’ and ‘Multicellular organismal process’ were the most significantly enriched term between both up and down-regulated isoforms.

I summarized the enriched KEGG pathways using DEIs. In blood, the most significantly enriched KEGG pathways in up-regulated isoforms were ‘Antigen processing and presentation’, ‘B cell receptor signaling pathway’, ‘Colorectal cancer’ and ‘MAPK signaling pathway’ but down-regulated isoforms were no enriched KEGG pathways. In skeletal muscle, the most significantly enriched KEGG pathways in down-regulated isoforms were ‘Axon guidance’, ‘Cell adhesion molecules (CAMs)’, ‘ECM−receptor interaction’ and ‘Glycosphingolipid biosynthesis’, the most significantly

enriched KEGG pathways in down-regulated isoforms were ‘Cytokine-cytokine receptor interaction’, ‘MAPK signaling pathway’, ‘Pathways in cancer’, ‘Jak-STAT signaling pathway’, ‘Toll-like receptor signaling pathway’, ‘Aldosterone-regulated sodium reabsorption’, ‘Pathogenic Escherichia coli infection’, ‘Bladder cancer’, ‘NOD-like receptor signaling pathway’, ‘Chemokine signaling pathway’, ‘Hematopoietic cell lineage’ and ‘p53 signaling pathway’ with the top 12 of the p-value (Figure 5.5)

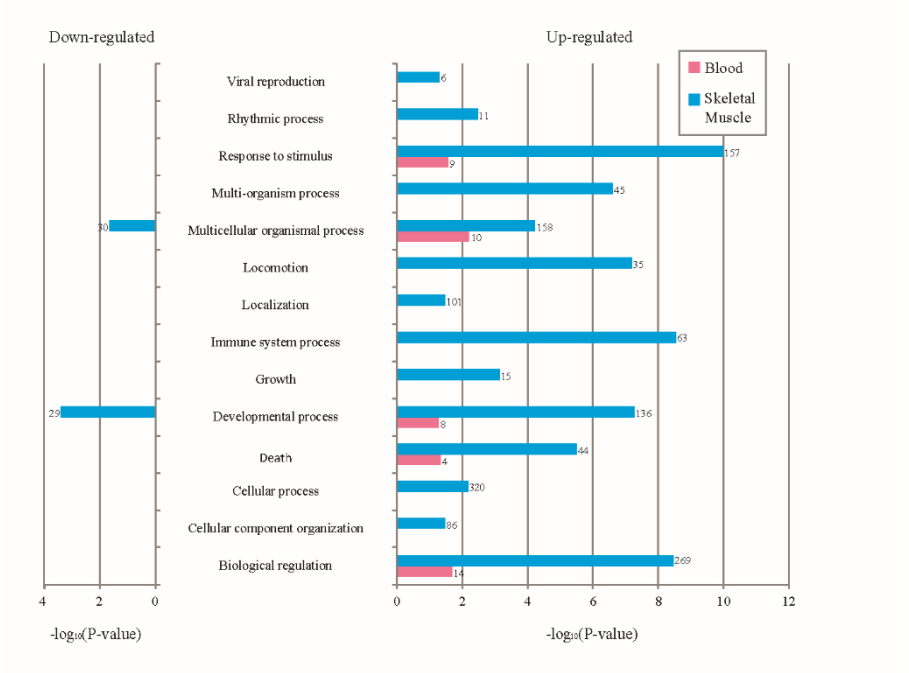


Figure 5.4. Highest category BP GO terms of tissues specific DEIs between before and after exercise in Thoroughbred.

Up–regulated isoforms indicate higher activation after exercise than before and down-regulation genes indicate lower activation after exercise than before exercise.

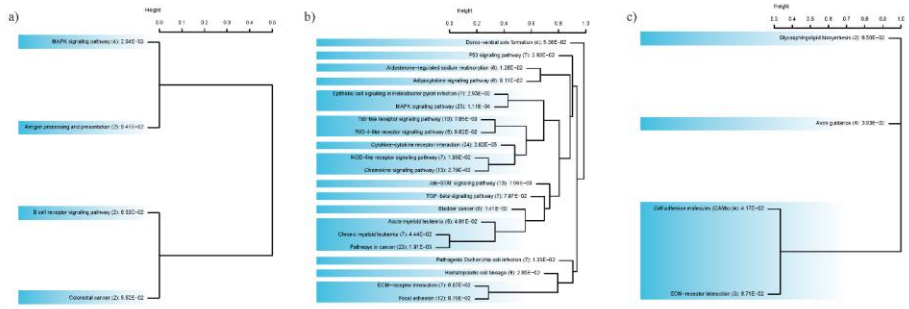


Figure 5.5. Hierarchical clustering of Enriched KEGG pathways associated with DEIs in two tissue such as blood and skeletal muscle

a) up-regulated DEIs in blood, b) up-regulated DEIs and c) down-regulated DEIs in skeletal muscle. The gene list of each Enriched KEGG pathways clustered using DAVID was compared to calculate the distance between the pathways. For a distance value >0.5, Enriched KEGG pathways were re-clustered, and pathways groups are shown as light-blue graduated blocks.

5.5 Discussion

5.5.1 Differentially expressed isoforms and splicing identification

In this study, blood and skeletal muscle RNA-seq data in six Thoroughbreds racing horses before and after exercise were publicized in more than 2 papers about DEGs with exercise-induce. The representative papers, at first, K.D.Park et al paper was publicized in BMC genomics in 2012 years, the second, Kim et al paper was publicized in DNA Research in 2013 and the third, W. Park et al was publicized in *PLOS One* in 2014. Among them, Kim et al and W. Park et al papers have something in common, at first, which were identified new concept genes that through the combination of evolution analysis and differential expressed analysis from Thoroughbred racing horse re-sequencing data and RNA-seq data respectively, the second, which were using edgeR program (Robinson et al. 2010) to find the DEGs however there was a very substantial difference that was applying horse reference when processed the assembly. Kim et al paper was applying but W.Park et al paper was not applying horse reference. Thoroughbred racing horse RNA-seq data were paired from the same horses in before and after exercise, so these 2 papers used the edgeR program that provide paired t-test but this program isn't provide the alternative splicing and differential splicing information that haven't been mentioned in Kim et al and W. Park et al papers. For these reason, I am strict examine the alternative splicing and differential splicing in Thoroughbred racing horse RNA-seq data. As a result, I identified that alternative splicing distribution, significant differential splicing and DEIs,

especially I except the DEIs that the intersect genes between DEGs and DEIs looking for transcripts which were differentially expressed in only isoforms. As a result, I identified 7 (6 up-regulated and 1 down-regulated) DEIs in blood and 56 (45 up-regulated and 11 down-regulated) DEIs in skeletal muscle. Among them, in blood, 3 isoforms such as *HSPA8*, *RhoB* and *SOCS3* (up-regulated) were revealed that closely related to exercise-induces. *HSPA8* gene was related to muscle hypertrophy in after 4 Hour exercise and it was up-regulated differentially expressed gene (McGivney et al. 2009). In other hands, this gene play an important role in rat cardiac muscle to protecting the heart against hypoxic exposure stress (Ikeda et al. 1999) and changed the gene expression in human neutrophils through aerobic exercise (Radom-Aizik et al. 2008). *RhoB* gene was closely related the Rho-kinase following activates Rho-kinase that help to the cooling strengthen contraction in cutaneous equine digital veins (Zerpa et al. 2010) and thereby increases RhoB expression in oxygen species (Kajimoto et al. 2007). *SOCS3* gene was reported that an important negative regulator of STAT3 activation and cytokine signaling following acute resistance exercise (Trenerry et al. 2008). In skeletal muscle, 5 isoforms such as *AMPD2, ICAM1, MMP-1, MXD1* and *TET2* were too revealed that related to exercise-induces. *AMPD2* gene was related to ATP following the major role of exercise-induces in human striated muscle, vigorous exercise condition, ATP level and *AMPD2* gene expression level were reported that seems to be a similar trend (Haas et al. 2003). *ICAM1* gene was increased by muscle overload with vigorous exercise and contributes to skeletal muscle hypertrophy as indicated by greater elevations in muscle mass, myofiber size, and protein content in mice (Dearth 2011). up-regulated *MMP*-

1 gene played an important role for the flow-enhanced motility in vascular smooth muscle cells (Shi et al. 2009). *MXD1* gene was a master regulator network with c-MYC as a vitamin D receptor (Salehi-Tabar et al. 2012), vitamin D deficiency caused muscle weakness (Holick 2007). For these reasons, I speculated that activates *MXD1* gene expression can cause increased vitamin D following the enhanced muscle. *TET2* gene in TET family was revealed that play a new role in DNA methylation and demethlation, active DNA methylation and demethlation remain to be determined that the play a part in the adaptive response of skeletal muscle to exercise training (Rasmussen et al. 2014).

Alternative splicing is play an important role in regulation of gene expression in eukaryotic species following the aspect of cell survival and function. However, alternative splicing studies in Thoroughbred racing horse were weakly understood. K.D.Park et al paper only mentioned that the gene expression pattern of alternative splicing existed in horse reference. Consequently, I identified the distribution of alternative splicing event type and differential splicing that was my main research purpose. I identified that exon skipping/inclusion (ESI) type is the most common of alternative splicing event in blood and skeletal muscle from Thoroughbred racing horse before and after exercise. This is identical result such as human and yeast (Sultan et al. 2008, Wang et al. 2008), in contrast to the report in pig (Chen et al. 2011). The percentage of ESI in total alternative splicing events in this study were 25.72% in blood and 27.27% in skeletal muscle, this percentage is lower than the reported human and rice (Zhang et al. 2010). I identified 4 differential splicing in skeletal muscle such as *BLZF1* (basic leucine zipper nuclear factor

1), *ITGB6* (Integrin, Beta 6), *KDM5C* (Lysine (K)-Specific Demethylase 5C) and *ZNF207* (zinc finger protein 207). Most of these genes have only one transcript, except for the *KDM5C* gene which have 2 transcript. Because I used -g/--GTF-guide option in cufflink program for founding novel gene, isoform, splicing. Therefore, I can speculate that novel transcript may exist in these genes. Although all genes have not been revealed that alternative splicing was directly related, I speculate that *ITGB6* and *ZNF207* genes were included a litter-related exercise-induce stress with alternative splicing because *ITGB6* gene was identified cell adhesion proteins that progression of oral squamous cell carcinoma is promoted by activating Fyn on binding to fibronectin, characterized lymph node metastasis (Li et al. 2003, Tamoto et al. 2004). In addition, this gene is play an important role in response to inflammation and high fat-diet (Qi et al. 2010). *ZNF207* gene was identified alternative splicing (inclusion of exon 9) in human breast cancer cells and a plays an role in apoptosis (Li et al. 2006). In Figure 5.2 and Figure 5.3, all of these genes have different pattern of read density in before and after exercise and alternative splicing events were existed between exon and exon.

5.5.2 Functional annotation of DEIs

The most GOBP term and KEGG pathway in my results look very similar to the results of Kim et al and W.Park et al paper. GOBP term in skeletal muscle, ‘Death’, ‘Immune system process’ and ‘Response to stimulus’ were related to immune response that was generated exercise-induced stress, as most know, muscle damage by exercise-induced was common phenomenon with

inflammatory response (Tidball 1995, Kingston et al. 1996, Clarkson et al. 1999, Stefansson et al. 2004). In addition, KEGG pathways such as ‘Cytokine-cytokine receptor interaction’, ‘MAPK signaling pathway’, ‘Jak-STAT signaling pathway’ and ‘p53 signaling pathway’ were also revealed to exercise-induced stress with inflammatory, immune responses, myogenesis, carbohydrate metabolism and regulation of apoptosis pathways (Long et al. 2004, Gourlay et al. 2005, Wang et al. 2008, Saleem et al. 2009). Though these results in GOBP term and KEGG pathway, I suggests that DEIs, in common with DEGs, also have enough ability to interpret.

General discussion

After NGS technology was invented, high-throughput technique for detection of biological meaning was dramatically developed. As a result, there are many methods and tools based on higher computational skills and transcriptome analysis in order to consider characteristics of the developed technique by handling massive data derived from RNA-seq. By using data acquired from NGS technology such as RNA-seq or additional DNA re-sequencing, lots of biological and evolutionary meaning could be obtained. In this thesis, I focused on the transcriptome analysis derived from RNA-seq and additional evolutionary analysis derived from DNA re-sequencing, especially in suitable choice of the transcriptome analysis workflows corresponding to the experimental design.

In chapter 2, the experimental design was simple comparison between adult Berkshire and jeju native pig using RNA-seq, this data was generated without replicates. So, RNA-seq analysis was implemented using DESeq tool that can detect DEGs without replicates. As a result, I identified 153 (87 up-regulated, 66 down-regulated), 169 (90 up-regulated, 79 down-regulated) and 39 (17 up-regulated, 22 down-regulated) DEGs in fat, liver and muscle, respectively, differentially expressed between JNP and Berkshire (FDR <0.01). In addition, Of the DEGs, 26 genes were related to meat quality and body growth, and functional annotation results were directly or indirectly related to meat quality and body growth. These results can be used as a valuable resource in future pig transcriptome analysis when there is no replicates and when simply two groups are compared. In addition, results

suggest that it is important for future RNA-seq studies to take the effect of breed into account.

In chapter 3, the experimental design was ordinal data and 3 replicates data in chicken broiler kidney under varied calcium intake. So, RNA-seq analysis was implemented using edgeR tool that is able to detect DEGs when RNA-seq data is ordinal and have replicates. Moreover, I implemented additional tool such as cuffdiff for simple two group comparison. As an experimental result, a decrease in the BW, BWG and FI, by high Ca intakes, could be observed. As a RNA-seq analysis results, I identified differentially expressed genes (DEGs) using cufflinks (128 DEGs between 0.8 and 1.0 percent, 141 DEGs between 0.8 and 1.2 percent and 103 DEGs between 1.0 and 1.2 percent), and also 12 DEGs were identified by edgeR. In summary, I demonstrate the empirical result that concentration of Ca increase leads to reduced BWG and FI by using transcriptome analysis such as the pathway enrichment, protein association networks and co-occurrence analysis from DEGs that are found by using methods such as cuffdiff and edgeR. First, I identified DEGs that directly are related to weight gain. Second, I also identified DEGs that are related to stress-induced disease such as hypertension that affects weight gain. These findings contribute to a better understanding of the molecular mechanisms potentially underlying correlation among Ca intakes, BWG, FI and stress-induced such as hypertension, and may provide important information relevant to other species, especially humans. In addition, these results can be used as a valuable resource in future chicken transcriptome analysis for comparison of two ordinal groups with replicates.

In chapter 4 and 5, RNA-seq data from six Thoroughbreds before and after exercise had already been published in DNA Research journal, and reference genome based RNA-seq analysis had been implemented using Thoroughbreds skeletal muscle and blood samples. The experimental design was paired sample (e.g. each of Thoroughbred has 4 sample such as before and after exercise samples in skeletal muscle and blood). For that reasons, I implemented de novo based RNA-seq analysis using Trinity tool. In addition, differentially expressed isoforms (DEIs), splicing and an alternative splicing event frequency were also identified using cufflinks tool. These studies were not performed in previous study done by reference genome-based RNA-seq analysis. As a results, I identified conceptually new DEGs involved in exercise response that have been selected during the domestication history of the Thoroughbred. Such a result cannot be acquired by reference based RNA-seq analysis. Moreover, 67 and 1,133 DEIs were identified in the blood and skeletal muscle respectively. 4 significant differential splicing were identified, and I identified that exon skipping/inclusion (ESI) type is the most common of alternative splicing event in blood and skeletal muscle from Thoroughbred racing horse before and after exercise. These results can be used as a valuable resource in future horse transcriptome analysis when paired sample comparison, de novo based RNA-seq analysis and additional analysis such as isoform and splicing. In addition, I suggest that it is important for future RNA-seq studies to take the application of de novo based RNA-seq analysis and other various RNA-seq analysis into account.

Thus, I suggest that researchers should be employ suitable transcriptome analysis corresponding to their experimental design.

References

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins and R. F. Galle (2000). The genome sequence of Drosophila melanogaster. Science **287**(5461): 2185-2195.

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde and R. F. Moreno (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252**(5013): 1651-1656.

Adams, M. D., A. R. Kerlavage, C. Fields and J. C. Venter (1993). 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nature genetics **4**(3): 256-267.

Adams, M. D., A. R. Kerlavage, J. M. Kelley, J. D. Gocayne, C. Fields, C. M. Fraser and J. C. Venter (1994). A model for high-throughput automated DNA sequencing and analysis core facilities. Nature **368**: 474-475.

Alterovitz, G. and M. F. Ramoni (2010). Knowledge based bioinformatics, Wiley Online Library.

Anders, S. (2010). Analysing RNA-Seq data with the DESeq package. Molecuar biology **43**(4): 1-17.

Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. Genome biology **11**(10): 1.

Andersson, L. (2012). How selective sweeps in domestic animals provide new insight into biological mechanisms. Journal of internal medicine **271**(1): 1-14.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Reference Source.

Anson, M., A.-M. Crain-Denoyelle, V. Baud, F. Chereau, A. Gougelet, B. Terris, S. Yamagoe, S. Colnot, M. Viguier and C. Perret (2012). Oncogenic β-catenin triggers an inflammatory response that determines the aggressiveness of hepatocellular carcinoma in mice. The journal of clinical investigation **122**(2): 586-599.

Applegate, T., R. Angel and H. Classen (2003). Effect of dietary calcium, 25-hydroxycholecalciferol, or bird strain on small intestinal phytase activity in broiler chickens. Poultry science **82**(7): 1140-1148.

Applegate, T. J. and R. Angel (2014). Nutrient requirements of poultry publication: History and need for an update. The journal of applied poultry research: japr980.

Auer, P. L. and R. W. Doerge (2011). A two-stage Poisson model for testing RNA-seq data. Statistical applications in genetics and molecular biology **10**(1).

Ayari, H. (2015). FABP4 Expression as Biomarker of Atheroma Development: A Mini-Review. Journal of molecular biomarkers & diagnosis **2015**.

Azain, M. (2004). Role of fatty acids in adipocyte growth and development. Journal of animal science **82**(3): 916-924.

Bîlek, K., A. Knoll, A. Stratil, K. Svobodová, P. Horák, R. Bechynova, M. Van Poucke and L. Peelman (2008). Analysis of mRNA expression of CNN3, DCN, FBN2, POSTN, SPARC and YWHAQ genes in porcine foetal and adult skeletal muscles. Czech journal of animal science **53**(5): 181.

Bagheri, R., A. N. Qasim, N. N. Mehta, K. Terembula, S. Kapoor, S. Braunstein, M. Schutta, N. Iqbal, M. Lehrke and M. P. Reilly (2010). Relation of plasma fatty acid binding proteins 4 and 5 with the metabolic syndrome, inflammation and coronary calcium in patients with type-2 diabetes mellitus. The american journal of cardiology **106**(8): 1118-1123.

Bainbridge, M. N., R. L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham and V. Magrini (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC genomics **7**(1): 1.

Ballard, P. L., J. W. Lee, X. Fang, C. Chapin, L. Allen, M. R. Segal, H. Fischer, B. Illek, L. W. Gonzales and V. Kolla (2010). Regulated gene expression in cultured type II cells of adult human lung. American journal of physiology-lung cellular and molecular Physiology **299**(1): L36-L50.

Barbazuk, W. B., S. J. Emrich, H. D. Chen, L. Li and P. S. Schnable (2007). SNP discovery via 454 transcriptome sequencing. The plant journal **51**(5): 910-918.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological): 289-300.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes and H. R. Bignell (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**(7218): 53-59.

Berry, D. C., D. DeSantis, H. Soltanian, C. M. Croniger and N. Noy (2012). Retinoic acid upregulates preadipocyte genes to block adipogenesis and suppress diet-induced obesity. Diabetes **61**(5): 1112-1121.

Birol, I., S. D. Jackman, C. B. Nielsen, J. Q. Qian, R. Varhol, G. Stazyk, R. D. Morin, Y. Zhao, M. Hirst and J. E. Schein (2009). De novo transcriptome assembly with ABySS. Bioinformatics **25**(21): 2872-2877.

Birzele, F., J. Schaub, W. Rust, C. Clemens, P. Baum, H. Kaufmann, A. Weith, T. W. Schulz and T. Hildebrandt (2010). Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucleic acids research **38**(12): 3999-4010.

Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode and G. F. Mayhew (1997). The complete genome sequence of Escherichia coli K-12. Science **277**(5331): 1453-1462.

Bloom, J. S., Z. Khan, L. Kruglyak, M. Singh and A. A. Caudy (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC genomics **10**(1): 221.

Boguski, M. S., T. M. Lowe and C. M. Tolstoshev (1993). dbEST—database for "expressed sequence tags. Nature genetics **4**(4): 332-333.

Bolger, A. M., M. Lohse and B. Usadel (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics: btu170.

Bolstad, B. M., R. A. Irizarry, M. Å strand and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**(2): 185-193.

Boon, R. A., C. Urbich, A. Fischer, R. D. Fontijn, F. H. Seeger, M. Koyanagi, A. J. Horrevoets and S. Dimmeler (2010). Krüppel-like factor 2 improves

neovascularization capacity of aged proangiogenic cells. European heart journal: ehq137.

Borch, K., C. Axelsson, H. Halgreen, M. D. Nielsen, T. Ledin and P. Szesci (1989). The ratio of pepsinogen A to pepsinogen C: a sensitive test for atrophic gastritis. Scandinavian journal of gastroenterology **24**(7): 870-876.

Bray, N. J., P. R. Buckland, M. J. Owen and M. C. O'Donovan (2003). Cis-acting variation in the expression of a high proportion of genes in human brain. Human genetics **113**(2): 149-153.

Brettschneider, J., N. Hartmann, V. Lehmensiek, H. Mogel, A. C. Ludolph and H. Tumani (2011). Cerebrospinal fluid markers of idiopathic intracranial hypertension: Is the renin-angiotensinogen system involved. Cephalalgia **31**(1): 116-121.

Browning, B. L. and Z. Yu (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. The american journal of human genetics **85**(6): 847-861.

Bullard, J. H., E. Purdom, K. D. Hansen and S. Dudoit (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics **11**(1): 94.

Burge, C. B., T. Tuschl and P. A. Sharp (1999). 20 Splicing of Precursors to mRNAs by the Spliceosomes. Cold spring harbor monograph archive **37**: 525-560.

Burset, M., I. Seledtsov and V. Solovyev (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic acids research **28**(21): 4364-4375.

Bushinsky, D. A. and R. D. Monk (1998). Calcium. The lancet **352**(9124): 306-311.

Cameron, N. (1990). Genetic and phenotypic parameters for carcass traits, meat and eating quality traits in pigs. Livestock production science **26**(2): 119-135.

Cao, H., J. Robinson, Z. Jiang, J. Melville, S. Golovan, M. Jones and A. Verrinder Gibbins (2004). A high-resolution radiation hybrid map of porcine chromosome 6. Animal genetics **35**(5): 367-378.

Carneiro, M. O., C. Russ, M. G. Ross, S. B. Gabriel, C. Nusbaum and M. A. DePristo (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC genomics **13**(1): 1.

Cecchetti, L., N. D. Tolley, N. Michetti, L. Bury, A. S. Weyrich and P. Gresele (2011). Megakaryocytes differentially sort mRNAs for matrix metalloproteinases and their inhibitors into platelets: a mechanism for regulating synthetic events. Blood **118**(7): 1903-1911.

Chang, Z., G. Li, J. Liu, Y. Zhang, C. Ashby, D. Liu, C. L. Cramer and X. Huang (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome biology **16**(1): 1.

Chau Long, Y., U. Widegren and J. R. Zierath (2004). Exercise-induced mitogen-activated protein kinase signalling in skeletal muscle. Proceedings of the nutrition society **63**(02): 227-232.

Chen, C., H. Ai, J. Ren, W. Li, P. Li, R. Qiao, J. Ouyang, M. Yang, J. Ma and L. Huang (2011). A global view of porcine transcriptome in three tissues from a full-sib pair with extreme phenotypes in growth and fat deposition by paired-end RNA sequencing. BMC genomics **12**(1): 448.

Chen, G., K. P. Yin, C. Wang and T. L. Shi (2011). De novo transcriptome assembly of RNA-Seq reads with different strategies. Science china life sciences **54**(12): 1129-1133.

Chen, Z., X. Gao, T. Lei, X. Chen, L. Zhou, A. Yu, P. Lei, R. Zhang, H. Long and Z. Yang (2011). Molecular characterization, expression and chromosomal localization of porcine PNPLA3 and PNPLA4. Biotechnology letters **33**(7): 1327-1337.

Cheng, I., M. E. Klingensmith, N. Chattopadhyay, O. Kifor, R. R. Butters, D. I. Soybel and E. M. Brown (1998). Identification and Localization of the Extracellular Calcium-Sensing Receptor in Human Breast 1. The journal of clinical endocrinology & metabolism **83**(2): 703-707.

Chikamoto, K., H. Misu, H. Takayama, A. Kikuchi, K.-a. Ishii, F. Lan, N. Takata, N. Tajima-Shirasaki, Y. Takeshita and H. Tsugane (2016). Rapid response of the steatosis-sensing hepatokine LECT2 during diet-induced weight cycling in mice. Biochemical and biophysical research communications.

Cho, I., H. Park, C. Yoo, G. Lee, H. Lim, J. Lee, E. Jung, M. Ko, J. Lee and J. Jeon (2011). QTL analysis of white blood cell, platelet and red blood cell-

related traits in an F2 intercross between Landrace and Korean native pigs. Animal genetics **42**(6): 621-626.

Choi, J.-W., H. Liu, D. K. Choi, T. S. Oh, R. Mukherjee and J. W. Yun (2012). Profiling of gender-specific rat plasma proteins associated with susceptibility or resistance to diet-induced obesity. Journal of proteomics **75**(4): 1386-1400.

Clarkson, P. M. and S. P. Sayers (1999). Etiology of exercise-induced muscle damage. Canadian journal of applied physiology **24**(3): 234-248.

Cohen, J. C., R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson and H. H. Hobbs (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science **305**(5685): 869-872.

Collett, S. R. (2012). Nutrition and wet litter problems in poultry. Animal feed science and technology **173**(1): 65-75.

Compeau, P. E., P. A. Pevzner and G. Tesler (2011). How to apply de Bruijn graphs to genome assembly. Nature biotechnology **29**(11): 987-991.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo and X. Zhang (2016). A survey of best practices for RNA-seq data analysis. Genome biology **17**(1): 1.

Consortium, S. (1998). Genome sequence of the nematode C. elegans A platform for investigating biology. Science **282**: 2012-2018.

Corino, C., J. Mourot, S. Magni, G. Pastorelli and F. Rosi (2002). Influence of dietary conjugated linoleic acid on growth, meat quality, lipogenesis, plasma leptin and physiological variables of lipid metabolism in rabbits. Journal of animal science **80**(4): 1020-1028.

Corino, C., G. Oriani, L. Pantaleo, G. Pastorelli and G. Salvatori (1999). Influence of dietary vitamin E supplementation on" heavy" pig carcass characteristics, meat quality, and vitamin E status. Journal of animal science **77**(7): 1755-1761.

D'Alessandro, A., P. G. Righetti and L. Zolla (2009). The red blood cell proteome and interactome: an update. Journal of proteome research **9**(1): 144-163.

Da Wei Huang, B. T. S. and R. A. Lempicki (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols **4**(1): 44-57.

Dai, M., R. C. Thompson, C. Maher, R. Contreras-Galindo, M. H. Kaplan, D. M. Markovitz, G. Omenn and F. Meng (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC genomics **11**(Suppl 4): S7.

Damlaj, M., R. Amre, P. Wong and J. How (2014). Hepatic ALECT-2 Amyloidosis Causing Portal Hypertension and Recurrent Variceal Bleeding. American journal of clinical pathology **141**(2): 288-291.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth and S. T. Sherry (2011). The variant call format and VCFtools. Bioinformatics **27**(15): 2156-2158.

Daniels, T. F. (2009). Progress Toward Genomically Optimized Beef: Cholesterol Transport Pathways and Lipid Homeostasis, Washington State University.

Darwin, C. (1868). The variation of animals and plants under domestication, O. Judd.

de Leeuw, N., T. Dijkhuizen, J. Y. Hehir-Kwa, N. P. Carter, L. Feuk, H. V. Firth, R. M. Kuhn, D. H. Ledbetter, C. L. Martin and C. van Ravenswaaij-Arts (2012). Diagnostic interpretation of array data using public databases and internet sources. Human mutation **33**(6): 930-940.

De Paepe, K. (2015). Comparison of methods for differential gene expression using RNA-seq data.

Dearth, C. L. (2011). Contributions of ICAM-1 to the immunobiology of skeletal muscle hypertrophy.

Dennis Jr, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome Biology **4**(5): P3.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas and M. Hanna (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics **43**(5): 491-498.

Di, Y., D. W. Schafer, J. S. Cumbie and J. H. Chang (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Statistical applications in genetics and molecular biology **10**(1).

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel and J. Estelle (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics **14**(6): 671-683.

Dior, U. P., L. Kogan, H. H. Chill, N. Eizenberg, A. Simon and A. Revel (2014). Emerging Roles of microRNA in the embryo–endometrium cross talk. Seminars in reproductive medicine, Thieme medical publishers.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**(1): 15-21.

Dong, C., A. Beecham, S. Slifer, L. Wang, S. H. Blanton, C. B. Wright, T. Rundek and R. L. Sacco (2010). Genomewide linkage and peakwide association analyses of carotid plaque in Caribbean Hispanics. Stroke **41**(12): 2750-2756.

Dong, C., A. Beecham, L. Wang, S. H. Blanton, T. Rundek and R. L. Sacco (2012). Follow-up association study of linkage regions reveals multiple candidate genes for carotid plaque in Dominicans. Atherosclerosis **223**(1): 177-183.

Dousset, E., J. Avela, M. Ishikawa, J. Kallio, S. Kuitunen, H. Kyrolainen, V. Linnamo and P. V. Komi (2007). Bimodal recovery pattern in human skeletal muscle induced by exhaustive stretch-shortening cycle exercise. Medicine and science in sports and exercise **39**(3): 453.

Doust, A. N., L. Lukens, K. M. Olsen, M. Mauro-Herrera, A. Meyer and K. Rogers (2014). Beyond the single gene: How epistasis and gene-by-environment effects influence crop domestication. Proceedings of the national academy of sciences **111**(17): 6178-6183.

Dunn, J. R., J. Reed, D. Du Plessis, E. Shaw, P. Reeves, A. Gee, P. Warnke and C. Walker (2006). Expression of ADAMTS-8, a secreted protease with antiangiogenic properties, is downregulated in brain tumours. British journal of cancer **94**(8): 1186-1193.

Emrich, S. J., W. B. Barbazuk, L. Li and P. S. Schnable (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome research **17**(1): 69-73.

Enquobahrie, D. A., M. Meller, K. Rice, B. M. Psaty, D. S. Siscovick and M. A. Williams (2008). Differential placental gene expression in preeclampsia. American journal of obstetrics and gynecology **199**(5): 566. e561-566. e511.

Estivill, X. and L. Armengol (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. Plos genetics **3**(10): e190.

Excoffier, L. and H. E. LISCHER (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources **10**(3): 564-567.

Fangauf, R., H. Vogt and W. Penner (1961). Studies of calcium tolerance in chickens. Arch. geflugelk. **25**: 82-86.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty and J. M. Merrick (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269**(5223): 496-512.

Fonseca, N. A., J. Marioni and A. Brazma (2014). RNA-seq gene profiling-a systematic empirical comparison. Plos one **9**(9): e107026.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork and C. Von Mering (2013). STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research **41**(D1): D808-D815.

Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton and J. M. Kelley (1995). The minimal gene complement of Mycoplasma genitalium. Science **270**(5235): 397-404.

Frazee, A. C., G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg and J. T. Leek (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. Nature biotechnology **33**(3): 243-246.

Friendly, M. (2002). Corrgrams. The american statistician **56**(4): 316-324.

Fukushima, N. and M. Fukayama (2007). Mucinous cystic neoplasms of the pancreas: pathology and molecular genetics. Journal of hepato-biliary-pancreatic surgery **14**(3): 238-242.

Furuhashi, M., S. Ishimura, H. Ota, M. Hayashi, T. Nishitani, M. Tanaka, H. Yoshida, K. Shimamoto, G. S. Hotamisligil and T. Miura (2011). Serum fatty

acid-binding protein 4 is a predictor of cardiovascular events in end-stage renal disease. Plos one **6**(11): e27356.

Furuhashi, M., T. Mita, N. Moniwa, K. Hoshina, S. Ishimura, T. Fuseya, Y. Watanabe, H. Yoshida, K. Shimamoto and T. Miura (2015). Angiotensin II receptor blockers decrease serum concentration of fatty acid-binding protein 4 in patients with hypertension. Hypertension research **38**(4): 252-259.

Fuseya, T., M. Furuhashi, S. Yuda, A. Muranaka, M. Kawamukai, T. Mita, S. Ishimura, Y. Watanabe, K. Hoshina and M. Tanaka (2014). Elevation of circulating fatty acid-binding protein 4 is independently associated with left ventricular diastolic dysfunction in a general population. Cardiovascular diabetology **13**(1): 1.

Gaffney, B. and E. Cunningham (1988). Estimation of genetic trend in racing performance of thoroughbred horses. Nature **332**(6166): 722-724.

Garber, M., M. G. Grabherr, M. Guttman and C. Trapnell (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nature methods **8**(6): 469-477.

Gilbert, W. (1978). Why genes in pieces. Nature **271**(5645): 501.

Gim, J., S. Won and T. Park (2016). Conditional estimation of local pooled dispersion parameter in small-sample RNA-Seq data improves differential expression test. Journal of bioinformatics and computational biology **14**(05): 1644006.

Giorgi, C., G. W. Yeo, M. E. Stone, D. B. Katz, C. Burge, G. Turrigiano and M. J. Moore (2007). The EJC factor eIF4AIII modulates synaptic strength and neuronal protein expression. Cell **130**(1): 179-191.

Gjerlaug-Enger, E., L. Aass, J. Ødegard and O. Vangen (2010). Genetic parameters of meat quality traits in two pig breeds measured by rapid methods. Animal **4**: 1832-1843.

Goff, J. P. and R. L. Horst (1994). Calcium salts for treating hypocalcemia: carrier effects, acid-base balance, and oral versus rectal administration. Journal of dairy science **77**(5): 1451-1456.

Goffeau, A., B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq and M. Johnston (1996). Life with 6000 genes. Science **274**(5287): 546-567.

Gogusev, J., P. Duchambon, B. Hory, M. Giovannini, Y. Goureau, E. Sarfati and T. B. Drüeke (1997). Depressed expression of calcium receptor in parathyroid gland tissue of patients with hyperparathyroidism. Kidney international **51**(1): 328-336.

Goh, D. L., A. Patel, G. H. Thomas, G. S. Salomons, D. S. Schor, C. Jakobs and M. T. Geraghty (2002). Characterization of the human gene encoding α-aminoadipate aminotransferase (AADAT). Molecular genetics and metabolism **76**(3): 172-180.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz and F. M. Lauro (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proceedings of the national academy of sciences **103**(30): 11240-11245.

Gordon, A. and G. Hannon (2010). Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit.

Gourlay, C. W. and K. R. Ayscough (2005). The actin cytoskeleton: a key regulator of apoptosis and ageing. Nature reviews molecular cell biology **6**(7): 583-589.

Grützmann, R., C. Pilarsky, E. Staub, A. O. Schmitt, M. Foerder, T. Specht, B. Hinzmann, E. Dahl, I. Alldinger and A. Rosenthal (2003). Systematic isolation of genes differentially expressed in normal and cancerous tissue of the pancreas. Pancreatology **3**(2): 169-178.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury and Q. Zeng (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology **29**(7): 644-652.

Grant, G. R., M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, C. J. Stoeckert, J. B. Hogenesch and E. A. Pierce (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics **27**(18): 2518-2528.

Graugnard, D., K. Moyes, E. Trevisi, M. Khan, D. Keisler, J. Drackley, G. Bertoni and J. Loor (2012). Liver lipid content and inflammometabolic indices in peripartal dairy cows are altered in response to prepartal energy intake and postpartal intramammary inflammatory challenge. Journal of dairy science.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. TRENDS in genetics **17**(2): 100-107.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai and M. H.-Y. Fritz (2010). A draft sequence of the Neandertal genome. Science **328**(5979): 710-722.

Gregg, C., J. Zhang, B. Weissbourd, S. Luo, G. P. Schroth, D. Haig and C. Dulac (2010). High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science **329**(5992): 643-648.

Groenendijk, B. C., K. Van der Heiden, B. P. Hierck and R. E. Poelmann (2007). The role of shear stress on ET-1, KLF2, and NOS-3 expression in the developing cardiovascular system of chicken embryos in a venous ligation model. Physiology **22**(6): 380-389.

Gunn, H. (1987). Muscle, bone and fat proportions and muscle distribution of Thoroughbreds and other horses.

Guo, Y., S. Zhao, Q. Sheng, M. Guo, B. Lehmann, J. Pietenpol, D. C. Samuels and Y. Shyr (2015). RNAseq by Total RNA Library Identifies Additional RNAs Compared to Poly (A) RNA Library. BioMed research international **2015**.

Haas, A. L. and R. L. Sabina (2003). N-terminal extensions of the human AMPD2 polypeptide influence ATP regulation of isoform L. Biochemical and biophysical research communications **305**(2): 421-427.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li and M. Lieber (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols **8**(8): 1494-1512.

Hardcastle, T. J. and K. A. Kelly (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC bioinformatics **11**(1): 422.

Henschel, R., P. M. Nista, M. Lieber, B. J. Haas, L.-S. Wu and R. D. LeDuc (2012). Trinity RNA-Seq assembler performance optimization. Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond, ACM.

Hillier, L., G. Lennon, M. Becker, M. F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello and W. Gish (1996). Generation and

analysis of 280,000 human expressed sequence tags. Genome research **6**(9): 807-828.

Holick, M. F. (2007). Vitamin D deficiency. New england journal of medicine **357**(3): 266-281.

Hong, D., A. Rhie, S.-S. Park, J. Lee, Y. S. Ju, S. Kim, S.-B. Yu, T. Bleazard, H.-S. Park and H. Rhee (2012). FX: an RNA-Seq analysis tool on the cloud. Bioinformatics **28**(5): 721-723.

Honma, F., K. Shio, K. Monoe, Y. Kanno, A. Takahashi, J. Yokokawa, H. Kobayashi, H. Watanabe, A. Irisawa and H. Ohira (2008). Primary biliary cirrhosis complicated by polymyositis and pulmonary hypertension. Internal medicine **47**(7): 667-669.

Hosack, D. A., G. ennis Jr, B. T. Sherman, H. C. Lane and R. A. Lempicki (2003). Identifying biological themes within lists of genes with EASE. Genome biology **4**(10): R70.

House, M. G., L. Kohlmeier, N. Chattopadhyay, O. Kifor, T. Yamaguchi, M. S. Leboff, J. Glowacki and E. M. Brown (1997). Expression of an Extracellular Calcium-Sensing Receptor in Human and Mouse Bone Marrow Cells. Journal of bone and mineral research **12**(12): 1959-1970.

Howard, B. E., Q. Hu, A. C. Babaoglu, M. Chandra, M. Borghi, X. Tan, L. He, H. Winter-Sederoff, W. Gassmann and P. Veronese (2013). High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. Plos one **8**(10): e74183.

Hurwitz, S., I. Plavnik, A. Shapiro, E. Wax and H. T. A. BAR (1995). Calcium Metabolism and Requirements of Chickens Are Affected by Growth1'2.

Hwang, H.-J., T. W. Jung, B.-H. Kim, H. C. Hong, J. A. Seo, S. G. Kim, N. H. Kim, K. M. Choi, D. S. Choi and S. H. Baik (2015). A dipeptidyl peptidase-IV inhibitor improves hepatic steatosis and insulin resistance by AMPK-dependent and JNK-dependent inhibition of LECT2 expression. Biochemical pharmacology **98**(1): 157-166.

Ikeda, T., K. Abe, A. Ota and T. Ikenoue (1999). Heat shock protein 70 and heat shock cognate protein 70 messenger ribonucleic acid induction in the brains, hearts, and livers of neonatal rats after hypoxic stress. American journal of obstetrics and gynecology **180**(2): 457-461.

Ishibashi, H., A. Komori, S. Shimoda, Y. M. Ambrosini, M. E. Gershwin and M. Nakamura (2011). Risk factors and prediction of long-term outcome in primary biliary cirrhosis. Internal medicine **50**(1): 1-10.

Jain, P., N. M. Krishnan and B. Panda (2013). Augmenting transcriptome assembly by combining de novo and genome-guided tools. PeerJ **1**: e133.

Jamart, C., N. Benoit, J.-M. Raymackers, H. J. Kim, C. K. Kim and M. Francaux (2012). Autophagy-related and autophagy-regulatory genes are induced in human muscle after ultraendurance exercise. European journal of applied physiology **112**(8): 3173-3177.

Jared, D. (1997). Guns, germs, and steel: the fates of human societies. NY: WW Norton & Company **14**.

Jeffcott, L., P. Rossdale, J. Freestone, C. Frank and P. TOWERS-CLARK (1982). An assessment of wastage in Thoroughbred racing from conception to 4 years of age. Equine veterinary journal **14**(3): 185-198.

Jin, S., C. Kim, Y. Song, W. Jang, Y. Kim, J. Yeo, J. Kim and K. Kang (2001). Physicochemical characteristics of longissimus muscle between the Korean native pig and Landrace. Korean Journal food science animal resour **21**: 142-148.

Johnson, J., N. Vatistas, L. Castro, T. Fischer, F. Pipers and D. Maye (2001). Field survey of the prevalence of gastric ulcers in Thoroughbred racehorses and on response to treatment of affected horses with omeprazole paste. Equine veterinary education **13**(4): 221-224.

Jung, W. Y., S. G. Kwon, M. Son, E. S. Cho, Y. Lee, J. H. Kim, B.-W. Kim, J. H. Hwang, T. W. Kim and H. C. Park (2012). RNA-Seq Approach for Genetic Improvement of Meat Quality in Pig and Evolutionary Insight into the Substrate Specificity of Animal Carbonyl Reductases. Plos one **7**(9): e42198.

Kajimoto, H., K. Hashimoto, S. N. Bonnet, A. Haromy, G. Harry, R. Moudgil, T. Nakanishi, I. Rebeyka, B. Thébaud and E. D. Michelakis (2007). Oxygen Activates the Rho/Rho-Kinase Pathway and Induces RhoB and ROCK-1 Expression in Human and Rabbit Ductus Arteriosus by Increasing Mitochondria-Derived Reactive Oxygen Species A Newly Recognized Mechanism for Sustaining Ductal Constriction. Circulation **115**(13): 1777-1788.

Kanitz, A., F. Gypas, A. J. Gruber, A. R. Gruber, G. Martin and M. Zavolan (2015). Comparative assessment of methods for the computational inference

of transcript isoform abundance from RNA-seq data. Genome biology **16**(1): 1.

Karasik, D. and M. Cohen-Zinder (2012). The genetic pleiotropy of musculoskeletal aging. Frontiers in physiology **3**.

Kasahara, T. and M. Kasahara (2000). Three aromatic amino acid residues critical for galactose transport in yeast Gal2 transporter. Journal of biological chemistry **275**(6): 4422-4428.

Katz, Y., E. T. Wang, E. M. Airoldi and C. B. Burge (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods **7**(12): 1009-1015.

Kayar, S., H. Hoppeler, S. Lindstedt, H. Claassen, J. Jones, B. Essen-Gustavsson and C. Taylor (1989). Total muscle mitochondrial volume in relation to aerobic capacity of horses and steers. Pflügers Archiv **413**(4): 343-347.

Kazemian, M., M. Ren, J. X. Lin, W. Liao, R. Spolski and W. J. Leonard (2015). Comprehensive assembly of novel transcripts from unmapped human RNA-Seq data and their association with cancer. Molecular systems biology **11**(8): 826.

Kestin, A. S., P. A. Ellis, M. R. Barnard, A. Errichetti, B. A. Rosner and A. D. Michelson (1993). Effect of strenuous exercise on platelet activation state and reactivity. Circulation **88**(4): 1502-1511.

Kijas, J. and L. Andersson (2001). A phylogenetic study of the origin of the domestic pig estimated from the near-complete mtDNA genome. Journal of molecular evolution **52**(3): 302-308.

Kim, D., B. Langmead and S. L. Salzberg (2015). "HISAT: a fast spliced aligner with low memory requirements." Nature methods **12**(4): 357-360.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology **14**(4): 1.

Kim, D., P. Seong, S. Cho, J. Kim, J. Lee, C. Jo and D. Lim (2009). Fatty acid composition and meat quality traits of organically reared Korean native black pigs. Livestock science **120**(1): 96-102.

Kim, H., T. Lee, W. Park, J. W. Lee, J. Kim, B.-Y. Lee, H. Ahn, S. Moon, S. Cho and K.-T. Do (2013). Peeling Back the Evolutionary Layers of Molecular

Mechanisms Responsive to Exercise-Stress in the Skeletal Muscle of the Racing Horse. DNA research **20**(3): 287-298.

Kim, M. Y., S. Lee, K. Van, T.-H. Kim, S.-C. Jeong, I.-Y. Choi, D.-S. Kim, Y.-S. Lee, D. Park and J. Ma (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proceedings of the national academy of sciences **107**(51): 22032-22037.

Kim, N.-K., J.-H. Lim, M.-J. Song, O.-H. Kim, B.-Y. Park, M.-J. Kim, I.-H. Hwang and C.-S. Lee (2008). Comparisons of longissimus muscle metabolic enzymes and muscle fiber types in Korean and western pig breeds. Meat science **78**(4): 455-460.

Kim, S.-S., J.-R. Kim, J.-K. Moon, B.-H. Choi, T.-H. Kim, K.-S. Kim, J.-J. Kim and C.-K. Lee (2009). Transcriptional alteration of p53 related processes as a key factor for skeletal muscle characteristics in Sus scrofa. Molecules and cells **28**(6): 565-573.

Kim, S. W., J. H. Jung, K. T. Do, K. S. Kim, C. H. Do, J. kyu Park, Y. K. Joo, T. S. Kim, B. H. Choi and T. H. Kim (2007). Investigation of Single Nucleotide Polymorphisms in Porcine Candidate Gene for Growth and Meat Quality Traits in the Berkshire Breed. Journal of life science **17**(12): 1622-1626.

Kim, T., K. Kim, B. Choi, D. Yoon, G. Jang, K. Lee, H. Chung, H. Lee, H. Park and J. Lee (2005). Genetic structure of pig breeds from Korea and China using microsatellite loci analysis. Journal of animal science **83**(10): 2255-2263.

Kingston, S. and L. Hoffman-Goetz (1996). Effect of environmental enrichment and housing density on immune system reactivity to acute exercise stress. Physiology & behavior **60**(1): 145-150.

Kinsella, M., O. Harismendy, M. Nakano, K. A. Frazer and V. Bafna (2011). Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics **27**(8): 1068-1075.

Klont, R., L. Brocks and G. Eikelenboom (1998). Muscle fibre type and meat quality. Meat science **49**: S219-S229.

Kodzius, R., M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai and M. Harbers (2006). CAGE: cap analysis of gene expression. Nature methods **3**(3): 211-222.

Komolka, K., E. Albrecht, K. Wimmers, J. J. Michal and S. Maak (2012). Molecular Heterogeneities of Adipose Depots-Potential Effects on Adi-pose-Muscle Cross-Talk in Humans, Mice and Farm Animals. journal of genomics **1**: 89-102.

Kouba, M., M. Enser, F. Whittington, G. Nute and J. Wood (2003). Effect of a high-linolenic acid diet on lipogenic enzyme activities, fatty acid composition, and meat quality in the growing pig. Journal of animal science **81**(8): 1967-1979.

Kvam, V. M., P. Liu and Y. Si (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American journal of botany **99**(2): 248-256.

Lan, F., H. Misu, K. Chikamoto, H. Takayama, A. Kikuchi, K. Mohri, N. Takata, H. Hayashi, N. Matsuzawa-Nagata and Y. Takeshita (2014). LECT2 functions as a hepatokine that links obesity to skeletal muscle insulin resistance. Diabetes **63**(5): 1649-1664.

Lan, J., M.-G. Lei, Y.-B. Zhang, J.-H. Wang, X.-T. Feng, D.-Q. Xu, J.-F. Gui and Y.-Z. Xiong (2009). Characterization of the porcine differentially expressed PDK4 gene and association with meat quality. Molecular biology reports **36**(7): 2003-2010.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle and W. FitzHugh (2001). Initial sequencing and analysis of the human genome. Nature **409**(6822): 860-921.

Langmead, B., K. D. Hansen and J. T. Leek (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome biology **11**(8): 1.

Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. Nature methods **9**(4): 357-359.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology **10**(3): 1.

Larson, G., K. Dobney, U. Albarella, M. Fang, E. Matisoo-Smith, J. Robins, S. Lowden, H. Finlayson, T. Brand and E. Willerslev (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science **307**(5715): 1618-1621.

Larson, G. and D. Q. Fuller (2014). The evolution of animal domestication. Annual review of ecology, evolution, and systematics **45**: 115-136.

Larson, G., D. R. Piperno, R. G. Allaby, M. D. Purugganan, L. Andersson, M. Arroyo-Kalin, L. Barton, C. C. Vigueira, T. Denham and K. Dobney (2014). Current perspectives and the future of domestication studies. Proceedings of the national academy of sciences **111**(17): 6139-6146.

Law, C. W., Y. Chen, W. Shi and G. K. Smyth (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology **15**(2): 1.

Lawrence, M., W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan and V. J. Carey (2013). Software for computing and annotating genomic ranges. Plos comput biology **9**(8): e1003118.

Lefebvre, V. and P. Smits (2005). Transcriptional control of chondrocyte fate and differentiation. Birth defects research part C: embryo today:reviews **75**(3): 200-212.

Leng, N., J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart and C. Kendziorski (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics **29**(8): 1035-1043.

Leon-Novelo, L., C. Fuentes and S. Emerson (2015). Bayesian Estimation of Negative Binomial Parameters with Applications to RNA-Seq Data. ArXiv preprint arXiv:1512.00475.

Levin, J. Z., M. F. Berger, X. Adiconis, P. Rogov, A. Melnikov, T. Fennell, C. Nusbaum, L. A. Garraway and A. Gnirke (2009). Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome biology **10**(10): 1.

Levine, R. J., J. C. Hauth, L. B. Curet, B. M. Sibai, P. M. Catalano, C. D. Morris, R. DerSimonian, J. R. Esterlitz, E. G. Raymond and D. E. Bild (1997). Trial of calcium to prevent preeclampsia. New england journal of medicine **337**(2): 69-77.

Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson and C. N. Dewey (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics **26**(4): 493-500.

Li, C., M. Kato, L. Shiue, J. E. Shively, M. Ares and R.-J. Lin (2006). Cell Type and Culture Condition–Dependent Alternative Splicing in Human Breast Cancer Cells Revealed by Splicing-Sensitive Microarrays. Cancer research **66**(4): 1990-1999.

Li, H. and R. Durbin (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics **26**(5): 589-595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). The sequence alignment/map format and SAMtools. Bioinformatics **25**(16): 2078-2079.

Li, J. and R. Tibshirani (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Statistical methods in medical research **22**(5): 519-536.

Li, R., Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian and J. Li (2010). Building the sequence map of the human pan-genome. Nature biotechnology **28**(1): 57-63.

Li, X., Y. Yang, Y. Hu, D. Dang, J. Regezi, B. L. Schmidt, A. Atakilit, B. Chen, D. Ellis and D. M. Ramos (2003). αvβ6-Fyn signaling promotes oral cancer progression. Journal of biological chemistry **278**(43): 41646-41653.

Liao, Y., G. K. Smyth and W. Shi (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research **41**(10): e108-e108.

Liao, Y., G. K. Smyth and W. Shi (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**(7): 923-930.

Lindholm-Perry, A. K., G. A. Rohrer, L. A. Kuehn, J. W. Keele, J. W. Holl, S. D. Shackelford, T. L. Wheeler and D. J. Nonneman (2010). Genomic regions associated with kyphosis in swine. BMC genetics **11**(1): 112.

Lindner, A. and A. Dingerkus (1993). Incidence of training failure among Thoroughbred horses at Cologne, Germany. Preventive veterinary medicine **16**(2): 85-94.

Lindner, R. and C. C. Friedel (2012). A comprehensive evaluation of alignment algorithms in the context of RNA-seq. Plos one **7**(12): e52403.

Livak, K. J. and T. D. Schmittgen (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2ΔΔCT Method. Methods **25**(4): 402-408.

Lo, Y. D. and R. W. Chiu (2009). Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. Clinical chemistry **55**(4): 607-608.

Long, Y. C., U. Widegren and J. R. Zierath (2004). Exercise-induced mitogen-activated protein kinase signalling in skeletal muscle. PROCEEDINGS-NUTRITION SOCIETY OF LONDON, Cambridge Univ Press.

Looft, C. (2013). Trancriptome Analysis of Testis and Liver for Androstenone by Using RNA Sequencing. Plant and Animal Genome XXI Conference, Plant and Animal Genome.

Lu, T., G. Lu, D. Fan, C. Zhu, W. Li, Q. Zhao, Q. Feng, Y. Zhao, Y. Guo and W. Li (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome research **20**(9): 1238-1249.

Lu, Y. (2013). Performance Comparison of Five RNA-seq Alignment Tools, New Jersey Institute of Technology, Department of Computer Science.

Mörlein, D., M. Lungershausen, K. Steinke, A. R. Sharifi and C. Knorr (2012). A single nucleotide polymorphism in the CYP2E1 gene promoter affects skatole content in backfat of boars of two commercial Duroc-sired crossbred populations. Meat science.

Maier, A., B. Völker, E. Boles and G. F. Fuhrmann (2002). Characterisation of glucose transport in Saccharomyces cerevisiae with plasma membrane vesicles (countertransport) and intact cells (initial uptake) with single Hxt1, Hxt2, Hxt3, Hxt4, Hxt6, Hxt7 or Gal2 transporters. FEMS yeast research **2**(4): 539-550.

Makeeva, O., A. Sleptsov, E. Kulish, O. Barbarash, A. Mazur, E. Prokhorchuk, N. Chekanov, V. Stepanov and V. Puzyrev (2015). Genomic Study of Cardiovascular Continuum Comorbidity. Acta naturae **7**(3): 89.

Makwana, O., N. M. King, L. Ahles, O. Selmin, H. L. Granzier and R. B. Runyan (2010). Exposure to low-dose trichloroethylene alters shear stress gene expression and function in the developing chick heart. Cardiovascular toxicology **10**(2): 100-107.

Malek, M., J. C. Dekkers, H. K. Lee, T. J. Baas, K. Prusa, E. Huff-Lonergan and M. F. Rothschild (2001). A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. Mammalian genome **12**(8): 637-645.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. Trends in genetics **24**(3): 133-141.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen and Z. Chen (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**(7057): 376-380.

Marra, M., L. Hillier, T. Kucaba, M. Allen, R. Barstead, C. Beck, A. Blistain, M. Bonaldo, Y. Bowers and L. Bowles (1999). An encyclopedia of mouse genes. Nature genetics **21**(2): 191-194.

Marrone, G., R. Maeso-Díaz, G. García-Cardena, J. G. Abraldes, J. C. García-Pagán, J. Bosch and J. Gracia-Sancho (2015). KLF2 exerts antifibrotic and vasoprotective effects in cirrhotic rat livers: behind the molecular mechanisms of statins. Gut **64**(9): 1434-1443.

Martin, J., V. M. Bruno, Z. Fang, X. Meng, M. Blow, T. Zhang, G. Sherlock, M. Snyder and Z. Wang (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC genomics **11**(1): 663.

Martin, J. A. and Z. Wang (2011). Next-generation transcriptome assembly. Nature reviews genetics.

Maxam, A. M. and W. Gilbert (1977). A new method for sequencing DNA. Proceedings of the national academy of sciences **74**(2): 560-564.

Mc Meekan, C. (1940). Growth and development in the pig, with special reference to carcass quality characters, Cambridge univ press.

McEllistrem, M. C. (2009). Genetic diversity of the pneumococcal capsule: implications for molecular-based serotyping. Future microbiology **4**(7): 857-865.

McGivney, B. A., S. S. Eivers, D. E. MacHugh, J. N. MacLeod, G. M. O'Gorman, S. D. Park, L. M. Katz and E. W. Hill (2009). Transcriptional adaptations following exercise in thoroughbred horse skeletal muscle

highlights molecular mechanisms that lead to muscle hypertrophy. BMC genomics **10**(1): 638.

McGivney, B. A., P. A. McGettigan, J. A. Browne, A. C. O. Evans, R. G. Fonseca, B. J. Loftus, A. Lohan, D. E. MacHugh, B. A. Murphy and L. M. Katz (2010). Characterization of the equine skeletal muscle transcriptome identifies novel functional responses to exercise training. BMC genomics **11**(1): 398.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel and M. Daly (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **20**(9): 1297-1303.

Medina, I., J. Tárraga, H. Martínez, S. Barrachina, M. Castillo, J. Paschall, J. Salavert-Torres, I. Blanquer-Espert, V. Hernández-García and E. S. Quintana-Ortí (2016). Highly sensitive and ultrafast read mapping for RNA-seq analysis. DNA research: dsv039.

Meitern, R., R. Andreson and P. Hõrak (2014). Profile of whole blood gene expression following immune stimulation in a wild passerine. BMC genomics **15**(1): 1.

Mercer, T. R., D. J. Gerhardt, M. E. Dinger, J. Crawford, C. Trapnell, J. A. Jeddeloh, J. S. Mattick and J. L. Rinn (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nature biotechnology **30**(1): 99-104.

Metpally, R. P. R., S. Nasser, I. Malenica, A. Courtright, E. Carlson, L. Ghaffari, S. Villa, W. Tembe and V. Keuren-Jensen (2013). Comparison of analysis tools for miRNA high throughput sequencing using nerve crush as a model. Frontiers in genetics **4**: 20.

Miller, W., D. I. Drautz, A. Ratan, B. Pusey, J. Qi, A. M. Lesk, L. P. Tomsho, M. D. Packard, F. Zhao and A. Sher (2008). Sequencing the nuclear genome of the extinct woolly mammoth. Nature **456**(7220): 387-390.

Mironov, A. A., J. W. Fickett and M. S. Gelfand (1999). Frequent alternative splicing of human genes. Genome research **9**(12): 1288-1293.

Moon, J. K., K. S. Kim, J. J. Kim, B. H. Choi, B. W. Cho, T. H. Kim and C. K. Lee (2009). Differentially expressed transcripts in adipose tissue between Korean native pig and Yorkshire breeds. Animal genetics **40**(1): 115-118.

Morin, R. D., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. Jones and M. A. Marra (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques **45**(1): 81.

Morin, R. D., N. A. Johnson, T. M. Severson, A. J. Mungall, J. An, R. Goya, J. E. Paul, M. Boyle, B. W. Woolcock and F. Kuchenbauer (2010). Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nature genetics **42**(2): 181-185.

Morris, C. D. and M. E. Reusser (1995). Calcium intake and blood pressure: epidemiology revisited. Seminars in nephrology.

Morrissy, A. S., R. D. Morin, A. Delaney, T. Zeng, H. McDonald, S. Jones, Y. Zhao, M. Hirst and M. A. Marra (2009). Next-generation tag sequencing for cancer gene expression profiling. Genome research **19**(10): 1825-1835.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods **5**(7): 621-628.

Morzel, M., C. Terlouw, C. Chambon, D. Micol and B. Picard (2008). Muscle proteome and meat eating qualities of Longissimus thoracis of Blonded Aquitaine young bulls: A central role of HSP27 isoforms. Meat science **78**(3): 297-304.

Mourot, J., M. Kouba and P. Peiniau (1995). Comparative study of< i> in vitro</i> lipogenesis in various adipose tissues in the growing domestic pig (Sus domesticus). Comparative biochemistry and physiology part B: biochemistry and molecular biology **111**(3): 379-384.

Muñoz, G., E. Alcázar, A. Fernández, C. Barragán, A. Carrasco, E. de Pedro, L. Silió, J. Sánchez and M. Rodríguez (2011). "Effects of porcine< i> MC4R</i> and< i> LEPR</i> polymorphisms, gender and Duroc sire line on economic traits in Duroc× Iberian crossbred pigs." Meat Science **88**(1): 169-173.

Murray, M., G. SCHUSSER, F. Pipers and S. J. Gross (1996). Factors associated with gastric lesions in Thoroughbred racehorses. Equine veterinary journal **28**(5): 368-374.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyder (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**(5881): 1344-1349.

Nakamura, M., Y. Takii, M. Ito, A. Komori, T. Yokoyama, Y. Shimizu-Yoshida, M. Koyabu, M. Matsuyama, T. Mori and T. Kamihira (2006). Increased expression of nuclear envelope gp210 antigen in small bile ducts in primary biliary cirrhosis. Journal of autoimmunity **26**(2): 138-145.

National Research Council, N. (2005). Mineral Tolerance of Animals: Second Revised Edition. The national academies press, USA.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee and E. E. Eichler (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature **461**(7261): 272-276.

Nielsen, K. L., A. L. Høgh and J. Emmersen (2006). DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic acids research **34**(19): e133-e133.

Niess, A., H. Dickhuth, H. Northoff and E. Fehrenbach (1999). Free radicals and oxidative stress in exercise--immunological aspects. Exercise immunology review **5**: 22.

Nookaew, I., M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén and J. Nielsen (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic acids research: gks804.

Novodvorsky, P. and T. Chico (2014). The role of the transcription factor KLF2 in vascular development and disease. Progress molecular biology translate science **124**: 155-188.

NRC. (1994). Nutrient requirements of poultry, National Research Council. National Academy Press Washington USA.

O'Reilly, E. L. (2016). Acute phase proteins and biomarkers for health in chickens, University of Glasgow.

O'Shea, J. J., M. Pesu, D. C. Borie and P. S. Changelian (2004). A new modality for immunosuppression: targeting the JAK/STAT pathway. Nature reviews drug discovery **3**(7): 555-564.

Olsen, K. M. and J. F. Wendel (2013). A bountiful harvest: genomic insights into crop domestication phenotypes. Annual review of plant biology **64**: 47-70.

Olsson, M., M. Ekblom, L. Fecker, M. Kurkinen and P. Ekblom (1999). cDNA cloning and embryonic expression of mouse nuclear pore membrane glycoprotein 210 mRNA. Kidney international **56**(3): 827-838.

Ota, H., M. Furuhashi, S. Ishimura, M. Koyama, Y. Okazaki, T. Mita, T. Fuseya, T. Yamashita, M. Tanaka and H. Yoshida (2012). Elevation of fatty acid-binding protein 4 is predisposed by family history of hypertension and contributes to blood pressure elevation. American journal of hypertension **25**(10): 1124-1130.

Ovejero, C., C. Cavard, A. Périanin, T. Hakvoort, J. Vermeulen, C. Godard, M. Fabre, P. Chafey, K. Suzuki and B. Romagnolo (2004). Identification of the leukocyte cell-derived chemotaxin 2 as a direct target gene of β-catenin in the liver. Hepatology **40**(1): 167-176.

Pan, Q., M. A. Bakowski, Q. Morris, W. Zhang, B. J. Frey, T. R. Hughes and B. J. Blencowe (2005). Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends in genetics **21**(2): 73-77.

Park, B., N. Kim, C. Lee and I. Hwang (2007). Effect of fiber type on postmortem proteolysis in longissimus muscle of Landrace and Korean native black pigs. Meat science **77**(4): 482-491.

Park, K. D., J. Park, J. Ko, B. C. Kim, H. S. Kim, K. Ahn, K. T. Do, H. Choi, H. M. Kim and S. Song (2012). Whole transcriptome analyses of six thoroughbred horses before and after exercise using RNA-Seq. BMC genomics **13**(1): 473.

Park, W., J. Kim, H. J. Kim, J. Choi, J.-W. Park, H.-W. Cho, B.-W. Kim, M. H. Park, T.-S. Shin and S.-K. Cho (2014). Investigation of De Novo Unique Differentially Expressed Genes Related to Evolution in Exercise Response during Domestication in Thoroughbred Race Horses. Plos one **9**(3): e91418.

Pearson, R., J. Fleetwood, S. Eaton, M. Crossley and S. Bao (2008). Krüppel-like transcription factors: a functional family. The international journal of biochemistry & cell biology **40**(10): 1996-2001.

Peng, Z., Y. Cheng, B. C.-M. Tan, L. Kang, Z. Tian, Y. Zhu, W. Zhang, Y. Liang, X. Hu and X. Tan (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nature biotechnology **30**(3): 253-260.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell and S. L. Salzberg (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology **33**(3): 290-295.

Petkov, S. G., H. Marks, T. Klein, R. S. Garcia, Y. Gao, H. Stunnenberg and P. Hyttel (2011). In vitro culture and characterization of putative porcine embryonic germ cells derived from domestic breeds and Yucatan mini pig embryos at Days 20–24 of gestation. Stem cell research **6**(3): 226-237.

Pimentel, H. J., N. Bray, S. Puente, P. Melsted and L. Pachter (2016). Differential analysis of RNA-Seq incorporating quantification uncertainty. BioRxiv: 058164.

Polyak, K. and G. J. Riggins (2001). Gene discovery using the serial analysis of gene expression technique: implications for cancer research. Journal of clinical oncology **19**(11): 2948-2958.

Poole, D. (2004). Current concepts of oxygen transport during exercise. Equine and comparative exercise physiology **1**(01): 5-22.

Power, M. L., R. P. Heaney, H. J. Kalkwarf, R. M. Pitkin, J. T. Repke, R. C. Tsang and J. Schulkin (1999). The role of calcium in health and disease. American journal of obstetrics and gynecology **181**(6): 1560-1569.

Prather, R. S. (2013). Pig genomics for biomedicine. Nature biotechnology **31**(2): 122-124.

Qi, L., M. C. Cornelis, P. Kraft, K. J. Stanya, W. L. Kao, J. S. Pankow, J. Dupuis, J. C. Florez, C. S. Fox and G. Paré (2010). Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Human molecular genetics **19**(13): 2706-2715.

Qin, B. and R. A. Anderson (2012). An extract of chokeberry attenuates weight gain and modulates insulin, adipogenic and inflammatory signalling pathways in epididymal adipose tissue of rats fed a fructose-rich diet. British journal of nutrition **108**(04): 581-587.

Quereda, C., L. Orte, J. Sabater, J. Navarro-Antolin, J. Villafruela and J. Ortuno (1996). Urinary calcium excretion in treated and untreated essential hypertension. Journal of the american society of nephrology **7**(7): 1058-1065.

Radom-Aizik, S., F. Zaldivar Jr, S.-Y. Leu, P. Galassetti and D. M. Cooper (2008). Effects of 30 min of aerobic exercise on gene expression in human neutrophils. Journal of applied physiology **104**(1): 236-243.

Ram, J. L., A. S. Karim, E. D. Sendler and I. Kato (2011). Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the Illumina sequencing platform. Systems biology in reproductive medicine **57**(3): 162-170.

Rao, S. R., M. Raju, M. Reddy and P. Pavani (2006). Interaction between dietary calcium and non-phytate phosphorus levels on growth, bone mineralization and mineral excretion in commercial broilers. Animal feed science and technology **131**(1): 135-150.

Rasmussen, M., J. R. Zierath and R. Barrès (2014). Dynamic epigenetic responses to muscle contraction. Drug discovery today.

Rauschecker, M. L., S. M. Cologna, P. Xekouki, N. Nilubol, R. D. Shamburek, M. Merino, P. S. Backlund, A. L. Yergey, E. Kebebew and J. E. Balow (2015). Clinical Case Report: LECT2-Associated Adrenal Amyloidosis. AACE clinical case reports **1**(1): e59-e67.

Rehfeldt, C. and G. Kuhn (2006). Consequences of birth weight for postnatal growth performance and carcass quality in pigs as related to myogenesis. Journal of animal science **84**(13 suppl): E113-E123.

Reinhardt, J. A., D. A. Baltrus, M. T. Nishimura, W. R. Jeck, C. D. Jones and J. L. Dangl (2009). De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. Genome research **19**(2): 294-305.

Reynolds, J., B. S. Weir and C. C. Cockerham (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105**(3): 767-779.

Riccardi, D., J. Park, W.-S. Lee, G. Gamba, E. M. Brown and S. C. Hebert (1995). Cloning and functional expression of a rat kidney extracellular calcium/polyvalent cation-sensing receptor. Proceedings of the National Academy of Sciences **92**(1): 131-135.

Risso, D., J. Ngai, T. P. Speed and S. Dudoit (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology **32**(9): 896-902.

Roberts, A., H. Pimentel, C. Trapnell and L. Pachter (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics **27**(17): 2325-2329.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada and J. Q. Qian (2010). De novo assembly and analysis of RNA-seq data. Nature methods **7**(11): 909-912.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1): 139-140.

Robinson, M. D. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology **11**(3): R25.

Robinson, M. D. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology **11**(3): 1.

Robinson, M. D. and G. K. Smyth (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics **23**(21): 2881-2887.

Robinson, M. D. and G. K. Smyth (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9**(2): 321-332.

Rodríguez, C. and C.-L. Flores (2000). Mutations in GAL2 or GAL4 alleviate catabolite repression produced by galactose in Saccharomyces cerevisiae. Enzyme and microbial technology **26**(9): 748-755.

Rogers, K. V., C. K. Dunn, S. C. Hebert and E. M. Brown (1997). Localization of calcium receptor mRNA in the adult rat central nervous system by in situ hybridization. Brain research **744**(1): 47-56.

Rustemeyer, S., W. Lamberson, D. Ledoux, K. Wells, K. Austin and K. Cammack (2011). Effects of dietary aflatoxin on the hepatic expression of apoptosis genes in growing barrows. Journal of animal science **89**(4): 916-925.

Ryle, A., F. Sanger, L. Smith and R. Kitai (1955). The disulphide bonds of insulin. Biochemical journal **60**(4): 541.

Sabeti, P., S. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. Mikkelsen, D. Altshuler and E. Lander (2006). Positive natural selection in the human lineage. Science **312**(5780): 1614-1620.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll and R. Gaudet (2007). Genome-wide detection and characterization of positive selection in human populations. Nature **449**(7164): 913-918.

Saleem, A., P. J. Adhihetty and D. A. Hood (2009). Role of p53 in mitochondrial biogenesis and apoptosis in skeletal muscle. Physiological genomics **37**(1): 58-66.

Salehi-Tabar, R., L. Nguyen-Yamamoto, L. E. Tavera-Mendoza, T. Quail, V. Dimitrov, B.-S. An, L. Glass, D. Goltzman and J. H. White (2012). Vitamin D receptor as a master regulator of the c-MYC/MXD1 network. Proceedings of the national academy of sciences **109**(46): 18827-18832.

Salonen, J. T., J.-M. Aalto, P. Uimari and M. Pirskanen (2007). Novel genes and markers in essential arterial hypertension, Google Patents.

Samborski, A., A. Graf, S. Krebs, B. Kessler and S. Bauersachs (2013). Deep Sequencing of the Porcine Endometrial Transcriptome on Day 14 of Pregnancy. Biology of reproduction.

Sandri, M., J. Lin, C. Handschin, W. Yang, Z. P. Arany, S. H. Lecker, A. L. Goldberg and B. M. Spiegelman (2006). PGC-1α protects skeletal muscle from atrophy by suppressing FoxO3 action and atrophy-specific gene transcription. Proceedings of the national academy of sciences **103**(44): 16260-16265.

Sanger, F. (1977). Nucleotide sequence of bacteriophage φX174 DNA. Nature.

Sanger, F. and A. R. Coulson (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology **94**(3): 441-448.

Sanger, F., S. Nicklen and A. Coulson (1991). DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology (Reading, Mass.) **24**: 104-108.

Sanger, F., S. Nicklen and A. R. Coulson (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences **74**(12): 5463-5467.

Sanger, F., E. Thompson and R. Kitai (1955). The amide groups of insulin. Biochemical journal **59**(3): 509.

Sanghera, D. K. and P. R. Blackett (2012). Type 2 diabetes genetics: beyond GWAS. Journal of diabetes & metabolism **3**(198).

Sato, N., N. Fukushima, A. Maitra, C. A. Iacobuzio-Donahue, N. T. van Heek, J. L. Cameron, C. J. Yeo, R. H. Hruban and M. Goggins (2004). Gene expression profiling identifies genes associated with invasive intraductal

papillary mucinous neoplasms of the pancreas. The american journal of pathology **164**(3): 903-914.

Schlottmann, I., M. Ehrhart-Bornstein, M. Wabitsch, S. Bornstein and V. Lamounier-Zepter (2014). Calcium-dependent release of adipocyte fatty acid binding protein from human adipocytes. International journal of obesity **38**(9): 1221-1227.

Schulz, M. H., D. R. Zerbino, M. Vingron and E. Birney (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics **28**(8): 1086-1092.

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. Nature **200**(8): 16-18.

Sebastian, S., S. Touchburn, E. Chavez and P. Lague (1996). Efficacy of supplemental microbial phytase at different dietary calcium levels on growth performance and mineral utilization of broiler chickens. Poultry science **75**(12): 1516-1523.

Selle, P. H., A. J. Cowieson and V. Ravindran (2009). Consequences of calcium interactions with phytate and phytase for poultry and pigs. Livestock science **124**(1): 126-141.

Serpell, J. (1989). Pet-keeping and animal domestication: a reappraisal. The walking larder: patterns of domestication, pastoralism, and predation: 10-21.

Seyednasrollah, F., A. Laiho and L. L. Elo (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics **16**(1): 59-70.

Shah, S. P., M. Köbel, J. Senz, R. D. Morin, B. A. Clarke, K. C. Wiegand, G. Leung, A. Zayed, E. Mehl and S. E. Kalloger (2009). Mutation of FOXL2 in granulosa-cell tumors of the ovary. New england journal of medicine **360**(26): 2719-2729.

Shah, S. P., R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany and J. Senz (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature **461**(7265): 809-813.

Shahid, S. U., K. W. Li, J. Acharya, J. A. Cooper, S. Hasnain and S. E. Humphries (2016). Effect of six type II diabetes susceptibility loci and an

FTO variant on obesity in Pakistani subjects. European journal of human genetics **24**(6): 903-910.

Shi-Zheng, G. and Z. Su-Mei (2009). Physiology, affecting factors and strategies for control of pig meat intramuscular fat. Recent patents on food, nutrition & agriculture **1**(1): 59-74.

Shi, Z.-D., X.-Y. Ji, H. Qazi and J. M. Tarbell (2009). Interstitial flow promotes vascular fibroblast, myofibroblast, and smooth muscle cell motility in 3-D collagen I via upregulation of MMP-1. American journal of physiology-heart and circulatory physiology **297**(4): H1225-H1234.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones and I. Birol (2009). ABySS: a parallel assembler for short read sequence data. Genome research **19**(6): 1117-1123.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. Kent and L. E. Hood (1985). Fluorescence detection in automated DNA sequence analysis. Nature **321**(6071): 674-679.

Smyth, G. K. (2005). Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor, Springer**:** 397-420.

Soneson, C., M. I. Love and M. D. Robinson (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000 Research **4**.

Srikanchai, T., E. Murani, K. Wimmers and S. Ponsuksili (2010). Four loci differentially expressed in muscle tissue depending on water-holding capacity are associated with meat quality in commercial pig herds. Molecular biology reports **37**(1): 595-601.

Srivastava, S. and L. Chen (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic acids research **38**(17): e170-e170.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic acids research **6**(7): 2601-2610.

Stefansson, S., M. Yepes, N. Gorlatova, D. E. Day, E. G. Moore, A. Zabaleta, G. A. McMahon and D. A. Lawrence (2004). Mutants of plasminogen activator inhibitor-1 designed to inhibit neutrophil elastase and cathepsin G

are more effective in vivo than their endogenous inhibi Journal of biological chemistry **279**(29): 29981-29987.

Steidl, C., S. P. Shah, B. W. Woolcock, L. Rui, M. Kawahara, P. Farinha, N. A. Johnson, Y. Zhao, A. Telenius and S. B. Neriah (2011). MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. Nature **471**(7338): 377-381.

STEVEN, L. and J. SALZBERG (2005). Beware of mis—assembled genomes. Bioinformatics **21**(4): 320-324.

Storz, J. F. (2005). INVITED REVIEW: Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular ecology **14**(3): 671-688.

Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov and D. Parkhomchuk (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321**(5891): 956-960.

Sun, L., K. Ma, H. Wang, F. Xiao, Y. Gao, W. Zhang, K. Wang, X. Gao, N. Ip and Z. Wu (2007). JAK1–STAT1–STAT3, a key pathway promoting proliferation and preventing premature differentiation of myoblasts. The journal of cell biology **179**(1): 129-138.

Sundquist, A., M. Ronaghi, H. Tang, P. Pevzner and S. Batzoglou (2007). Whole-genome sequencing and assembly with high-throughput, short-read technologies. Plos one **2**(5): e484.

Surget-Groba, Y. and J. I. Montoya-Burgos (2010). Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome research **20**(10): 1432-1440.

Sutton, G. G., O. White, M. D. Adams and A. R. Kerlavage (1995). TIGR Assembler: A new tool for assembling large shotgun sequencing projects. Genome science and technology **1**(1): 9-19.

Tagliaro, C. H., M. H. L. P. Franco, W. Meincke and G. Silva (1995). Protein phenotypes and productive traits in landrace, large white and duroc swine. Ciência rural **25**(1): 61-65.

Takahashi, T., Y. Kasashima and Y. Ueno (2004). Association between race history and risk of superficial digital flexor tendon injury in Thoroughbred

racehorses. Journal of the american veterinary medical association **225**(1): 90-93.

Tamoto, E., M. Tada, K. Murakawa, M. Takada, G. Shindo, K.-i. Teramoto, A. Matsunaga, K. Komuro, M. Kanai and A. Kawakami (2004). Gene-expression profile changes correlated with tumor progression and lymph node metastasis in esophageal cancer. Clinical cancer research **10**(11): 3629-3638.

Tao, T. (2006). Program Parameters for blastall.

Tarazona, S., P. Furió-Tarí, D. Turrà, A. Di Pietro, M. J. Nueda, A. Ferrer and A. Conesa (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic acids research: gkv711.

Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). Differential expression in RNA-seq: a matter of depth. Genome research **21**(12): 2213-2223.

Tesfaye, B., S. Wuehler, T. Moges, A. Samuel, A. Kebede, D. Zerfu, A. Abera, G. Mengistu, B. Wodajo and K. A. Birks (2015). Making a Case for Calcium Supplementation for Prevention of Pregnancy Hypertension in Ethiopia.

Teshima, K. M., G. Coop and M. Przeworski (2006). How reliable are empirical genomic scans for selective sweeps. Genome research **16**(6): 702-712.

Tesseraud, S., R. Peresson, J. Lopes and A. Chagneau (1996). Dietary lysine deficiency greatly affects muscle and liver protein turnover in growing chickens. British journal of nutrition **75**(06): 853-865.

Thomas, R. K., E. Nickerson, J. F. Simons, P. A. Jänne, T. Tengs, Y. Yuza, L. A. Garraway, T. LaFramboise, J. C. Lee and K. Shah (2006). Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nature medicine **12**(7): 852-855.

Tidball, J. G. (1995). Inflammatory cell response to acute muscle injury. Medicine and science in sports and exercise **27**(7): 1022-1032.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology **31**(1): 46-53.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**(9): 1105-1111.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols **7**(3): 562-578.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology **28**(5): 511-515.

Trenerry, M. K., K. A. Carey, A. C. Ward, M. M. Farnfield and D. Cameron-Smith (2008). Exercise-induced activation of STAT3 signaling is increased with age. Rejuvenation research **11**(4): 717-724.

Trut, L. (1999). Early Canid Domestication: The Farm-Fox Experiment Foxes bred for tamability in a 40-year experiment exhibit remarkable transformations that suggest an interplay between behavioral genetics and development. American scientist **87**(2): 160-169.

Trut, L., I. Oskina and A. Kharlamova (2009). Animal evolution during domestication: the domesticated fox as a model. Bioessays **31**(3): 349-360.

Tso, A. W., A. Xu, P. C. Sham, N. M. Wat, Y. Wang, C. H. Fong, B. M. Cheung, E. D. Janus and K. S. Lam (2007). Serum Adipocyte Fatty Acid–Binding Protein as a New Biomarker Predicting the Development of Type 2 Diabetes A 10-year prospective study in a Chinese cohort. Diabetes care **30**(10): 2667-2672.

Tyra, M. and K. Ropka-Molik (2011). "Effect of the FABP3 and LEPR gene polymorphisms and expression levels on intramuscular fat (IMF) content and fat cover degree in pigs. Livestock science **142**(1): 114-120.

Uddin, M. J., D. N. Duy, M. U. Cinar, D. Tesfaye, E. Tholen, H. Juengst, C. Looft and K. Schellander (2011). Detection of quantitative trait loci affecting serum cholesterol, LDL, HDL, and triglyceride in pigs. BMC genetics **12**(1): 62.

Urban, T., R. Mikolasova, J. Kuciel, M. Ernst and I. Ingr (2002). A study of associations of the H-FABP genotypes with fat and meat production of pigs. Journal of applied genetics **43**(4): 505-510.

Van De Wiel, M. A., G. G. Leday, L. Pardo, H. Rue, A. W. Van Der Vaart and W. N. Van Wieringen (2012). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics: kxs031.

Van Verk, M. C., R. Hickman, C. M. Pieterse and S. C. Van Wees (2013). RNA-Seq: revelation of the messengers. Trends in plant science **18**(4): 175-179.

Vasa, M., S. Fichtlscherer, A. Aicher, K. Adler, C. Urbich, H. Martin, A. M. Zeiher and S. Dimmeler (2001). Number and migratory activity of circulating endothelial progenitor cells inversely correlate with risk factors for coronary artery disease. Circulation research **89**(1): e1-e7.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). Serial analysis of gene expression. Science **270**(5235): 484.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans and R. A. Holt (2001). The sequence of the human genome. Science **291**(5507): 1304-1351.

Vigne, J.-D. (2011). The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere." Comptes rendus biologies **334**(3): 171-181.

Visvader, J. E. and G. J. Lindeman (2003). Transcriptional regulators in mammary gland development and cancer. The international journal of biochemistry & cell biology **35**(7): 1034-1051.

Vitting-Seerup, K., B. T. Porse, A. Sandelin and J. Waage (2014). spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC bioinformatics **15**(1): 81.

Wagenknecht, D., H. Bartenschlager, M. Van Poucke, H. Geldermann, L. Peelman, I. Majzlik and A. Stratil (2005). Linkage and radiation hybrid mapping of the porcine MPZ gene to chromosome 4q. Animal genetics **36**(2): 181-182.

Wain, L. V. (2014). Blood pressure genetics and hypertension: genome-wide analysis and role of ancestry. Current genetic medicine reports **2**(1): 13-22.

Wain, L. V., G. C. Verwoert, P. F. O'Reilly, G. Shi, T. Johnson, A. D. Johnson, M. Bochud, K. M. Rice, P. Henneman and A. V. Smith (2011). Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nature genetics **43**(10): 1005-1011.

Walk, C., E. Addo-Chidie, M. Bedford and O. Adeola (2012). Evaluation of a highly soluble calcium source and phytase in the diets of broiler chickens. Poultry science **91**(9): 2255-2263.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge (2008). Alternative isoform regulation in human tissue transcriptomes. Nature **456**(7221): 470-476.

Wang, K., C. Wang, F. Xiao, H. Wang and Z. Wu (2008). JAK2/STAT2/STAT3 are required for myogenic differentiation. Journal of biological chemistry **283**(49): 34029-34036.

Wang, L., Z. Feng, X. Wang, X. Wang and X. Zhang (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics **26**(1): 136-138.

Wang, Y., S. Kim and I.-m. Kim (2014). Regulation of metastasis by microRNAs in ovarian cancer. Frontiers in oncology **4**: 143.

Wang, Y., S. Wang and Y. Zhang (2015). Leukocyte chemotactic factor 2 associated renal amyloidosis: one case report. Beijing da xue xue bao. Yi xue ban= Journal of peking university. Health sciences **47**(2): 349-351.

Warner, L. E., M. J. Hilz, S. H. Appel, J. M. Killian, E. H. Kolodny, G. Karpati, S. Carpenter, G. V. Watters, C. Wheeler and D. Witt (1996). Clinical Phenotypes of Different Mutations May Include Charcot–Marie–Tooth Type 1B, Dejerine–Sottas, and Congenital Hypomyelination. Neuron **17**(3): 451-460.

Weatherby, J. (1791). An Introduction to a General Stud Book. Weatherby and Sons, London.

Weber, A. P., K. L. Weber, K. Carr, C. Wilkerson and J. B. Ohlrogge (2007). Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant physiology **144**(1): 32-42.

White, A., M. Estrada, K. Walker, P. Wisnia, G. Filgueira, F. Valdés, O. Araneda, C. Behn and R. Martínez (2001). Role of exercise and ascorbate on plasma antioxidant capacity in thoroughbred race horses. Comparative biochemistry and physiology part A: molecular & integrative physiology **128**(1): 99-104.

Wilhelm, B. T. and J.-R. Landry (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods **48**(3): 249-257.

Winkelmann, R., L. Sandrock, M. Porstner, E. Roth, M. Mathews, E. Hobeika, M. Reth, M. L. Kahn, W. Schuh and H.-M. Jäck (2011). B cell homeostasis

and plasma cell homing controlled by Krüppel-like factor 2. Proceedings of the national academy of sciences **108**(2): 710-715.

Wood, J., R. Richardson, G. Nute, A. Fisher, M. Campo, E. Kasapidou, P. Sheard and M. Enser (2004). Effects of fatty acids on meat quality: a review. Meat science **66**(1): 21-32.

Wright, S. (1949). The genetical structure of populations. Annals of eugenics **15**(1): 323-354.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, M. D. McMullen and B. S. Gaut (2005). The effects of artificial selection on the maize genome. Science **308**(5726): 1310-1314.

Wu, J. Y., L. Yuan and N. Havlioglu (2004). Alternatively spliced genes. Encyclopedia of molecular cell biology and molecular medicine.

Xia, Z., H. Xu, J. Zhai, D. Li, H. Luo, C. He and X. Huang (2011). RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis. Plant molecular biology **77**(3): 299-308.

Xie, Y., G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu and S. Li (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics **30**(12): 1660-1666.

Yamagoe, S., Y. Yamakawa, Y. Matsuo, J. Minowada, S. Mizuno and K. Suzuki (1996). Purification and primary amino acid sequence of a novel neutrophil chemotactic factor LECT2. Immunology letters **52**(1): 9-13.

Yamamoto, M., T. Wakatsuki, A. Hada and A. Ryo (2001). Use of serial analysis of gene expression (SAGE) technology. Journal of immunological methods **250**(1): 45-66.

Yang, R., G. Castriota, Y. Chen, M. Cleary, K. Ellsworth, M. Shin, J.-L. Tran, T. Vogt, M. Wu and S. Xu (2011). RNAi-mediated germline knockdown of FABP4 increases body weight but does not improve the deranged nutrient metabolism of diet-induced obese mice. International journal of obesity **35**(2): 217-225.

Young, L., D. Marlin, C. Deaton, H. Brown-Feltner, C. Roberts and J. Wood (2002). Heart size estimated by echocardiography correlates with maximal oxygen uptake. Equine veterinary journal **34**(S34): 467-471.

Yu, X., T. Riley and A. J. Levine (2009). The regulation of the endosomal compartment by p53 the tumor suppressor gene. FEBS journal **276**(8): 2201-2212.

Zeder, M. A. (2012). The domestication of animals. Journal of anthropological research: 161-190.

Zeder, M. A. (2015). Core questions in domestication research. Proceedings of the national academy of sciences **112**(11): 3191-3198.

Zerbino, D. R. and E. Birney (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research **18**(5): 821-829.

Zerpa, H., Y. Berhane, H. Woodcock, J. Elliott and S. R. Bailey (2010). Rho kinase activation and ROS production contributes to the cooling enhanced contraction in cutaneous equine digital veins. Journal of applied physiology **109**(1): 11-18.

Zhai, W., L. Araujo, S. Burgess, A. Cooksey, K. Pendarvis, Y. Mercier and A. Corzo (2012). Protein expression in pectoral skeletal muscle of chickens as influenced by dietary methionine1. Poultry science **91**(10): 2548-2555.

Zhang, F. and R. Drabier (2012). IPAD: the integrated pathway analysis database for systematic enrichment analysis. BMC bioinformatics **13**(15): 1.

Zhang, G., G. Guo, X. Hu, Y. Zhang, Q. Li, R. Li, R. Zhuang, Z. Lu, Z. He and X. Fang (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome research **20**(5): 646-654.

Zhang, W. (2009). Involvement of protein degradation, calpain autolysis and protein nitrosylation in fresh meat quality during early postmortem refrigerated storage.

Zhao, J., J. J. Brault, A. Schild, P. Cao, M. Sandri, S. Schiaffino, S. H. Lecker and A. L. Goldberg (2007). FoxO3 coordinately activates protein degradation by the autophagic/lysosomal and proteasomal pathways in atrophying muscle cells. Cell metabolism **6**(6): 472-483.

Zhou, H., N. Deeb, C. Evock-Clover, C. Ashwell and S. Lamont (2006). Genome-wide linkage analysis to identify chromosomal regions affecting phenotypic traits in the chicken. II. Body composition. Poultry science **85**(10): 1712-1721.

Zhu, Y., M. Li, A. M. Sousa and N. Šestan (2014). XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. BMC genomics **15**(1): 1.

Zweig, A. S., D. Karolchik, R. M. Kuhn, D. Haussler and W. J. Kent (2008). UCSC genome browser tutorial. Genomics **92**(2): 75-84.

요약(국문초록)

다양한 실험 디자인으로부터 유래된 가축화 동물들의 RNA 시퀀싱 기반의 전사체 분석

박원철

농생명공학부 동물생명공학전공

서울대학교 대학원 농업생명과학대학

오늘날의 차세대 시퀀싱(NGS)은 한번의 수행으로 수십억 염기서열의 시퀀싱 데이터를 생산할 수 있다. 게다가 다양한 연구 분야에서 수천 개의 논문이 NGS 를 사용하여 연구되고 출판되고 있다. NGS 기술은 현재 생물과학과 진화과학에서 가장 영향력 있는 장비이고 지금까지 생산된 정보를 합친 것보다 더 많은 정보를 생산해 내고 있다. RNA-seq 은 차세대 시퀀싱이 발명 된 이후로 현재까지 이용되고 있는 최신 기술이다. 유전자 발현 프로파일 연구에서, 전사체 시퀀싱은 모든 전사체의 프로파일이 가능하기

때문에 가장 적합한 기술이라고 할 수 있다. 주어진 생물학적 시점에서 전 방위적 세포 전사 프로파일과 현저하게 향상된 RNA 검출 방법의 성능은 전체 전사체 시퀀싱에 의해 제공된다. RNA 를 위한 NGS 기술의 이용에서, 몇몇 연구는 성공적으로 수행 되었다. 가까운 미래에, 모든 연구자들은 RNA 시퀀싱과 같은 RNA 분야의 NGS 기술을 일상적으로 사용할 것 이다. 하지만 전사체 분석은 그 분야의 사람들이 쓰기에는 결코 쉽지만은 않다. 따라서 이 학위 논문은 복잡한 데이터인 유전자의 발현 정보나 진화 정보와 같은 NGS 데이터, 즉 RNA 시퀀싱 데이터를 이용한 연구들이 주가 되어 구성하였다.

제 1 장에서는 NGS 의 일반적인 배경 지식을 요약하였다. 먼저 NGS 기술의 역사 그리고 방법의 분류를 기술하였고, 조금 더 구체적으로 NGS 방법을 유전체와 전사체로 나누어 목록화 하였다. RNA-seq 의 특성도 요약하였다. 먼저 시퀀싱의 역사 그리고 유전자 발현에 대해 기술하였고 RNA-seq 과 이전의 연구들을 비교하였다. 그리고 RNA-seq 분석에 대한 전반적인 개요를 설명하였다. 마지막으로 가축화 유전자 (말, 돼지 그리고 닭)의 진화에 대해 소개하였다.

제 2 장에서는 제주 재래돼지 한 마리와 버크셔 한 마리의 서로 다른 3 조직(간, 지방 그리고 근육)의 RNA-seq 데이터를 이용하여, 각각의 조직에서 품종간의 반응으로 인해 유전자 발현 패턴의 유의한 변화를 조사하였다. 제주 재래 돼지는 제주도라는

특이한 환경에 적응되었으며, 질병 저항성으로 잘 알려져 있다. 특히 서구 품종 보다 육질이 부드러우며 육즙이 많고 채도가 높은 면에서 육질이 좋다고 알려져 있다. 이런 제주 재래 돼지의 특이한 표현형의 분자 메커니즘을 이해하기 위해서, 나는 RNA-seq 기술을 이용한 비교 전사체 연구를 시행하였다. 제주 재래 돼지와 버크셔의 3 조직을 서로 비교 하였으며 두 품종의 조직간 차등 발현 유전자를 찾아 냈다. 이 차등 발현 유전자 중에 나는 26 개의 유전자가 육질과 체중 증가에 연관성이 있다는 것을 밝혀냈다. 결과적으로 나는 제주 재래돼지가 버크셔에 비해 육질과 체중 증가와 관련이 있는 다른 유전자 발현 프로파일을 가지고 있다고 제안한다.

제 3 장에서는 닭의 육계 품종인 브로일러 9 마리를 3 마리씩 3 가지 칼슘 섭취량 그룹으로 나눈 RNA-seq 데이터를 이용하여 신장에서 칼슘 스트레스의 반응으로 유전자 발현 패턴의 유의한 변화를 조사하였다. 닭은 적색야계 라고 불리는 야생 종으로부터 처음 가축화 되었다. 적색야계는 아직도 남아시아에서 대부분 야생 종으로 활동 중이다. 그 이후 회색야계가 아마도 약 8000 년 전에 교잡되어 발생되었다. 그리고 가축화 된 닭은 경제 형질 아이디어에 의해 육계와 산란계로 현재까지 선택되어 길러지고 있다. 이 품종 중에 육계인 브로일러는 닭 산업에서 대부분의 비중을 차지하고 있다. 게다가 칼슘은 정상 세포기능 과 혈류의 응고에 필수적인 요소이다. 하지만 칼슘 적정 섭취량 보다 부족하거나 과한 경우에 저 칼슘혈증 및 고 칼슘혈증을 일으키기도 한다. 이와 같은 증상은

체중과 고혈압과 같은 스트레스와 연관성이 있다. 그래서 나는 칼슘 섭취량(0.8%, 1.0% 그리고 1.2%)에 따른 체중의 변화를 실험적 기법을 통해 실험 하였으며 이에 따른 RNA-seq 데이터도 생산 하였다. 그 결과 0.8% 과 1.0% 사이에서 123 개의 차등 발현 유전자를, 0.8%와 1.2% 사이에서 141 개의 차등 발현 유전자를, 그리고 1.0% 와 1.2% 사이에서 103 개의 차등 발현 유전자를 cufflinks 방법으로 밝혀냈다. 더욱이 12 개의 차등 발현 유전자를 edgeR 의 순차적 방법으로 밝혀냈다. KEGG pathway enrichment, the co-occurrence 와 the protein/protein interaction (PPI) 네트워크 분석을 통해 연구한 결과, 이 유전자들은 혈압과 고혈압과 관련이 있었다. 그 이후 7 개의 차등 발현 유전자를 무작위로 선택하여 qRT-PRC 방법으로 검증을 하였다. 요약하면, 이 논문의 목적은 칼슘 섭취량의 변화가 육계에서 신장에서 어떤 영향을 미치는지 조사하는 것이며, 결과적으로 나는 적정량 보다 높은 칼슘 섭취량은 브로일러에서 체중감소와 체중감소에 영향을 미치는 고혈압과 같은 스트레스 유도 질병을 유발 한다고 제안 한다.

제 4 장에서는 이전의 연구에서 분석한 참조 유전체 기반 RNA-seq 분석이 아닌, *de novo* 기반 RNA-seq 분석을 통해 6 마리의 더러브렛 경주마의 운동 전 후에 생산된 RNA-seq 데이터를 이용하여 근육과 혈액에서 운동 스트레스에 반응하는 전 유전체적 발현 패턴을 분석하였다. 나는 2 가지 주된 아이디어에 집중하였다. 첫 번째로, *de novo* 기반 RNA-seq 분석을 통해 차등 발현 유전자를 밝혀내는데 이는 이전 연구에서 밝혀내지 못한 차등

발현 유전자만을 밝혀내는데 목적이 있다. 두 번째로, 더러브렛과 제주 포니의 유전체 re-sequencing 데이터를 이용하여 가축화 유전자를 찾고 이와 차등 발현 유전자를 결합해서 새로운 개념의 차등 발현 유전자를 밝혀내는 것이 목적이다. 그 결과 나는 근육에서 1,034 개의 차등발현 유전자, 혈액에서 567 개의 차등 발현 유전자를 찾았다. 이 유전자들은 이 전 연구에서는 차등 발현 유전자라고 발견이 되지 않은 유전자였다. 근육에서의 차등 발현 유전자들은 운동에 의한 스트레스에 유의하게 반응한 염증 반응과 세포 자멸사에 관련되어 있었다. 더욱이 이 연구에서, 전사체 분석과 진화 분석을 통해 획득한 5 개의 운동과 관련된 유전자들은 더러브렛의 진화 역사에서 가축화 선택 유전자 이기도 하다고 밝혔으면 이는 새로운 개념의 차등 발현 유전자라고 설명할 수 있다.

제 5 장에서는 4 장과 마찬가지로 이전 6 마리의 더러브렛 경주마의 운동 전 후에 생산된 RNA-seq 데이터를 바탕으로 기존에 분석하지 않은 다양한 분석을 하였다. 특히 차등 발현 동형 단백질, differentially splicing 그리고 대체 짜집기 이벤트 (alternative splicing event) 을 cufflink 방법을 통해 조사하였다. 그 결과 근육에서 1.133 개의 차등 발현 동형 단백질, 혈액에서 67 개의 차등 발현 동형 단백질을 찾을 수 있었다. 그리고 이 들은 선행 연구와 마찬가지로 운동 스트레스에 반응 하는 차등 발현 동형 단백질이며, 염증 반응과 세포 자멸사에 관련이 있었다. 그리고 4 개의 유의한 differential splicing 을 밝혔으며 이들 중

2 개만 운동 스트레스와 관련이 있었다. 그리고 나는 대체 짜집기 이벤트 중에 exon skipping/inclusion (ESI)가 더러브렛에서 더 많이 일어 난다는 것을 밝혔으며 이는 인간과 효모와도 같은 현상이었다. 하지만 돼지에서 alternative 3’ splicing (A3)과는 다른 결과를 나타냈다.

이러한 연구들을 통해, 실험 디자인과 목적을 고려해서 전사체 분석의 다양한 이용은 NGS 로부터 유래된 RNA-seq 데이터 또는 추가적인 DNA re-sequencing 데이터에서 성공적으로 논증되었다. RNA-seq 와 추가적인 DNA re-sequencing 기술로부터 얻은 정보를 사용하여, 많은 생물학적 진화학적 의미들을 얻을 수 있었다. 주어진 이 결과들로, 나는 전사체 연구 분야에서 종사하는 연구자들은 실험적 디자인 및 목적에 적합한 전사체 분석을 사용해야 한다고 제한하는 바이다.

주요어: 차세대 시퀀싱, 가축화 동물, 전사체 분석, 차등 발현 유전자, 드노보 어셈블리, 차등 발현 동형 단백질, 선택적 이어 맞추기.

학번: 2009-21249